MOLECULAR APPROACHES TO ELUCIDATING ADAPTATION TO SERPENTINE SOILS USING THE *CAULANTHUS AMPLEXICAULIS* COMPLEX (BRASSICACEAE)

A Dissertation

by

ANGELA KRISTA HAWKINS

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Alan Pepper
Committee Members,	Lisa Campbell
	Jessica Light
	Thomas McKnight
Head of Department,	Thomas McKnight

August 2018

Major Subject: Biology

ABSTRACT

Serpentine endemic plants are excellent models for the study of molecular evolution as they provide extreme examples of adaptation to environment. Serpentine outcrops are derived from ultramafic rock and have low levels of essential plant nutrients (e.g., N, P, K, Ca), as well as toxic levels of heavy metals (e.g., Ni, Cr, Co), very poor moisture availability, high levels of light, and elevated soil temperatures. These outcrops provide habitat to the endemic plant species, *Caulanthus amplexicaulis* var. *barbarae* (CAB). Its sister species, *C. amplexicaulis* var. *amplexicaulis* (CAA), is found predominately on granite soils and is intolerant to serpentine soils.

Comprehensive reference transcriptomes of CAA and CAB were assembled and annotated for use in protein coding gene comparisons. Orthologs between CAA and CAB reveal high genome-wide d*N*/d*S* ratios and result from the composite effects of drift, positive selection, and the relaxation of negative selection. Also, paralogs within each taxon revealed two periods of elevated gene duplication. Further, distribution of d*S* is strongly bimodal indicating two distinct divergence events between the taxa, and suggesting that introgression may have contributed to serpentine adaptation.

Common-garden and reciprocal transplant experiments were performed on natural granite and serpentine soils using CAA and CAB. RNA-seq analyses were implemented to calculate global expression patterns and identify differentially regulated genes that may play a role in serpentine adaptation. Initial efforts were implemented to answer the following three questions: which genes are constitutively expressed in CAB, which genes are induced in CAB on serpentine outcrops, and which genes are induced in CAA on serpentine outcrops? RNA-seq data implicates

a suite of chloroplast and plastids related genes being constitutively expressed in CAB; this is an unexpected and novel finding. Genes induced in CAB on serpentine outcrops include those with roles in nutrient acquisition and transport and heavy metal binding. Genes induced in CAA on serpentine outcrops indicate response to nutrient starvation and galactose binding/transport. Ultimately results from these analyses, in conjunction with QTL mapping and population genetic data, will be used to find high quality candidate genes that confer tolerance to serpentine soils.

DEDICATION

I dedicate this to my husband, best friend, and co-author, Wm. Daryl Hawkins. Without his constant support and encouragement, I would not be writing this.

ACKNOWLEDGEMENTS

I would like to thank my committee chair, Dr. Alan Pepper, for taking a risk and allowing me to join his lab. He is single handedly the most patient human being I have ever known and his calmness, intelligence, and understanding has kept me in check. I also thank my committee members for their time, help, and critiques - all of which has made me a better scientist. Dr. McKnight has offered up some really great ideas to make this work better and he is one of the funniest people I know. Thanks for teaching me 112 and making those pickles. Dr. Light (an animal parasitologist agreeing to being on a plant person's committee nonetheless) has helped me run several programs when I was at my wit's end and allowed me to present plant papers at her journal club within the wildlife and fisheries department. And to Dr. Campbell who was always excited to listen to my research and offer great advice based on her own student's work and who's prelim exam made me learn more about some aspects of my research than I thought possible.

I would also like to offer my gratitude to the department's staff, who in my humble opinion, are the best group of people and keep this ship sailing. Anita was always there to help me with Concur so I just didn't give up and not get the refund. Will who always made sure I got paid, sometimes even too much! Lieu is the most responsive person on earth and would always help me reserve a room or answer silly questions. And to Jennifer, who's the best graduate student advisor, ever. As was stated in a letter for one of her much deserved nominations: she was a friend when you needed it, a mom when you needed it, and a warden when you needed to get back on track (i.e. have a committee meeting).

Of course I need to thank my graduate student family that I have been lucky to have been a part of for the past *cough* 8 years. No matter what is going on in your life, one of them is going through the same thing. Big thanks to Sachi and Wangming for, among other things, help with using bioinformatics programs. Sarah Beagle, Adam Foxfire, Drew Anderson, Chloe Bennett, Patrick Suess thank you for your friendship, laughs, and teaching me how to make those dots on this table of contents. Former grads Sarah Herlihy and Emily Rose for literally making me laugh until I cried. My favorite pastime is still watching the campus police pull over bikes. And of course I thank my lab mate for the past 5 years, Elyssa Garza. She's helped me with everything from data analysis to talking me out of a panic attack at 10,000 feet.

For the rest of my life I will always thank my master's advisor, Dr. Chris Randle, for allowing an old, mediocre grade having undergrad to be, not only his grad student, but his first one. That could have gone downhill quickly. It didn't.

And of course I thank the most important person in my life, my husband Daryl. Without him I would not even have a bachelor's degree. He has constantly been supportive financially, emotionally, and unconditionally. He's my best friend and the love of my life. He is literally the smartest person I know and he has a heart of gold.

I will end with this: achieving this honor of having a PhD in a field that I have loved since before I knew the word biology has been the second most exciting thing to happen to me this year. The first came on the evening of February 4th, 2018 when my hometown Philadelphia Eagles pulled off the greatest upset in the history of sports and won the Super Bowl. Even better, they beat the Patriots to do so. So I thank each one of the world champion Eagles, the past 4 months would have been a LOT less fun had they lost. Fly Eagles Fly!

CONTRIBUTORS AND FUNDING SOURCES

This work was supported by a dissertation committee consisting of Dr. Alan Pepper [advisor] and Dr. Thomas McKnight of the Department of the Biology, Dr. Jessica Light of the Department of Wildlife and Fisheries, and Dr. Lisa Campbell who has dual appointments with the Biology and Oceanography departments.

The data for figure 8 in Chapter II were obtained by Valerie Dietz, an undergraduate in the Pepper lab. This work was published, with Dietz as a co-author, in Genome Biology and Evolution in December, 2017. The data analyzed for Chapter III was provided from growth and RNA-seq experiments performed by Dr. Alan Pepper. The WGCNA analyses explained in Chapter III were conducted in part by Sachi Mandal of the Biology Department. All other work conducted for the dissertation was completed by the student independently.

Graduate study was supported by a teaching assistantship from Texas A&M University and, in part, from a research grant funded by NSF Foundation (IOS 12581020).

NOMENCLATURE

Ath	*Arabidopsis thaliana*
CAA	*Caulanthus amplexicaulis* var. *amplexicaulis*
CAB	*Caulanthus amplexicaulis* var. *barbarae*
d*N*	non-synonymous substitution rate
d*S*	synonymous substitution rate
d*N*/d*S*	non-synonymous to synonymous ratio
GO	Gene ontology term
MS	Murashige and Skoog medium
MA	Million years ago
N	Population size
N_e	Effective population size
PAML	Phylogenetic Analysis by Maximum Likelihood
RBH	Reciprocal best BLAST hit
RIL	Recombinant inbred line
RTL	Representative transcript locus
TAIR	The *Arabidopsis* Information Resource
TOP(s)	Tentative orthologous pair(s)
TPM	Transcripts per million
TPPs	Tentative paralogous pair(s)
WGCNA	Weighted correlation network analysis
ω_S	Synthetic d*N*/d*S*

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

CHAPTER I

INTRODUCTION AND LITERATURE REVIEW

"Nothing can be more abrupt than the change often due to the diversity of soil."

-Alfred Wallace (1858)

In 1995, Niles Eldredge stated that "adaptation is the heart and soul of evolution". Plants occupy some of the most arduous environment conditions, undoubtedly exhibiting some of the most sophisticated adaptations known. Depending on their specific environment, plants face many diverse and unique challenges including exposure to extreme temperatures, competition for sunlight and other resources, and herbivory, as well as challenges induced from soil diversity. While it is naïve to suggest, *a priori*, that all adaptations have evolved via natural selection for their current function, Wallace suggested that the ability of plants to adapt to different soil types is usually indicative of strong natural selection caused by ecological variation (Wallace, 1858).

Serpentine soils are derived from the alteration or "serpentinizaton" of ultramafic rock which contains at least 70% of mafic minerals such as magnesium and iron silicates (Brooks, 1983; Brady, 2005). These soils are found on every continent and in every type of ecosystem; however, they are usually patchily distributed. Serpentine soils have three uniting features: 1) poor productivity from flora and fauna, 2) high rates of endemism, and 3) unique vegetation types compared to those found in adjacent areas (Whittaker, 1954). Serpentine soils have long been considered an ideal model system for plant molecular ecology. Several naturally occurring examples of conspecifics have shown extreme habitat preference (e.g., serpentine endemic versus granite outcrop), allowing highly informative reciprocal transplant experiments to be performed. Further, serpentine specific phenotypes can easily be observed in laboratory settings

and these plants are amenable to manipulation and growth experiments in greenhouse settings. Finally, whole plant physiology has been described opening the path for understanding the genetic basis for adaptation to serpentine soils (Brady, 2005).

Roughly 1.5% (6000 km^2) of California soil is considered serpentine and in that area alone, ~13% of all Californian endemic flora are found. *Caulanthus amplexicaulis* var. *barbarae* (CAB), is a small annual diploid that is endemic to a series of isolated serpentine outcrops in the San Rafael Mountains of southwest California while its sister taxon, *C. amplexicaulis* var. *amplexicaulis* (CAA), has a much wider distribution and is mainly found on granite soils throughout the Transverse Ranges of southern California (Pepper & Norwood, 2001). Although CAB and CAA are ecologically and geographically isolated (at their closest they are found separated by 75 km), they are morphologically similar with only sepal color as an obvious identifier (Fig. 1). However, CAB and CAA are fully interfertile in artificial crosses (Kruckeberg, 1984; Pepper & Norwood, 2001). CAB and CAA are easy to work with in laboratory settings and have a generation time (seed to seed) of apporximately 10-12 weeks. Further, both parentals and hybrid offspring can be selfed for several generations easily resulting in recombinant inbred lines (RILs).

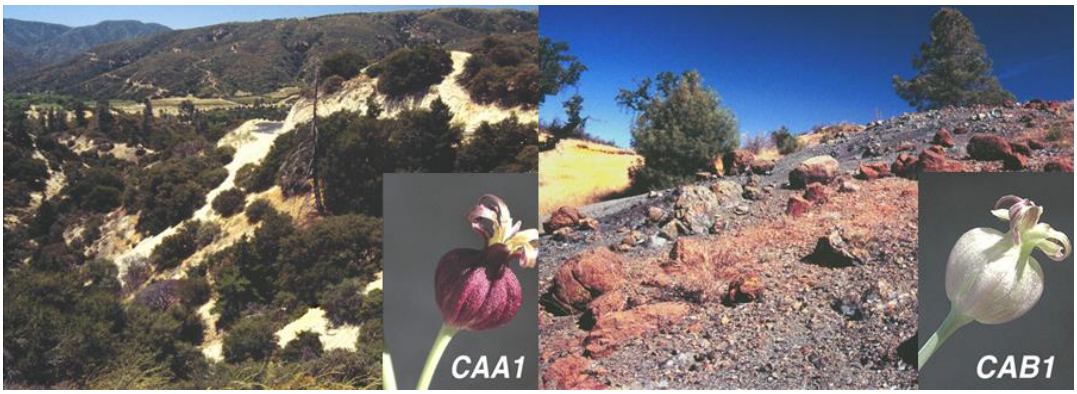

Figure 1. CAA and CAB morphological and edaphic variation.

CAB and CAA have recently been redefined as members of the Strepthanoid Clade I (SC-I; Cacho et al., 2014), which includes other serpentine endemic species as well as non-serpentine members. The genome sizes of CAB and CAA are estimated to be about 372 megabase pairs (MBP) (Johnston et al., 2005), roughly 2.5 times that of *Arabidopsis thaliana,* have a conserved chromosome count (n=14), and contain 25,000 – 29,500 genes (Pepper, unpublished data). The Department of Energy Joint Genome Institute has sequenced the CAB genome and Pepper lab member Elyssa Garza has assembled and annotated both CAB and CAA genomes. These data in conjunction with transcriptome data, RNA-seq data (described within), and QTL data will be used to determine the molecular basis for adaption to serpentine soils.

Serpentine plants share the ability to overcome three major challenges: 1) a low calcium to magnesium ratio, 2) low to absent levels of essential plant nutrients, and 3) high to lethal levels of toxic metals (Kazakou et al., 2008). These well-documented observations have led to the fundamental questions that this research proposes to answer:

1. What are the most critical environmental challenges that must be overcome to achieve tolerance to serpentine soils? A long-standing list of hypotheses has attempted to explain serpentine tolerance. Kruckeberg (1984) warns against pointing to any single, universal explanatory cause, and suggests that several factors work in concordance to produce what he and Jenny (1980) termed the "serpentine syndrome". To determine the key ecological limitation(s) that CAB must adapt to the following 'abnormal' edaphic conditions have been tested in the Pepper lab: 1) absent or deficient levels of essential plant nutrients (such as N, P, K, Ca, S), 2) high to lethal amounts of toxic, heavy metals (such as Ni, Cu, Cd, Zn), 3) other environmental stresses (such as salinity, high light, and temperature).

2. What are the molecular mechanisms that underlie adaptation to the serpentine environment? Fundamental growth experiments have provided substantial evidence of phenotypic differences between CAA and CAB environments and set the foundation for subsequent molecular work that will help elucidate the genetic basis for adaptation to serpentine soils. Molecular analyses will help determine, among other things, whether serpentine tolerance stems from unique genes/gene products or differential expression of certain genes.

To begin testing which environmental challenges are most important for allowing or prohibiting growth on serpentine soils, preliminary laboratory growth and phenotypic experiments were performed by several members of the Pepper lab. Seeds from inbred parental lines of CAA and CAB were planted in 'ideal' environments: those emulating the soils and conditions under which both species occur naturally, and in a varying degree of stressful environments. Conetainers™ (part no. RLC4 pine, Stuewe & Sons) were first filled half way with perlite, then topped with a mixture of high quality silica glass sand (U.S. Silica, Texas Coarse no. 2) and acrylamide soil water retention beads (Aquadiamonds Soil Polymers®). To provide the germinating seeds a constant form of moisture, a $1cm^2$ cube of rock wool was placed inside the top layer of sand. One seed was planted per conetainer, 200 conetainers per holding rack (part no. RL200, Stuewe & Sons), and placed in enclosed plastic tubs to monitor and manipulate humidity conditions for the duration of the growth experiments. All seeds were hydrated with purified water until cotyledons were visible and opened, after which varying nutrient treatments (to saturation) of plants began and continued every 48 hours for the duration of the experiment.

Murashige and Skoog (MS) medium (MSP01), ¼ strength, with salts, micro-, and macronutrients, pH 5.8 (Murashige & Skoog, 1962) was used to grow plants in as close to an

optimal nutrient solution as possible. To recreate environmental conditions that CAB plants are exposed to, several independent growth experiments were undertaken. Plants were stressed by treating with: 1) low phosphorus, 2) low nitrogen, 3) high nickel, and 4) high levels of magnesium compared to calcium. All seeds were uniformly germinated and treatment began when the first true leaves (cotyledons) emerged. Plants were grown for six-weeks then harvested and allowed to dry completely before weighing. Biomass accumulation was used as a proxy for overall fitness; the control experiment (using complete medium for both variations) showed no difference between CAA and CAB plants; however, CAB significantly ($P = 0.0001$) outperformed CAA in every stressed treatment (Fig. 2).

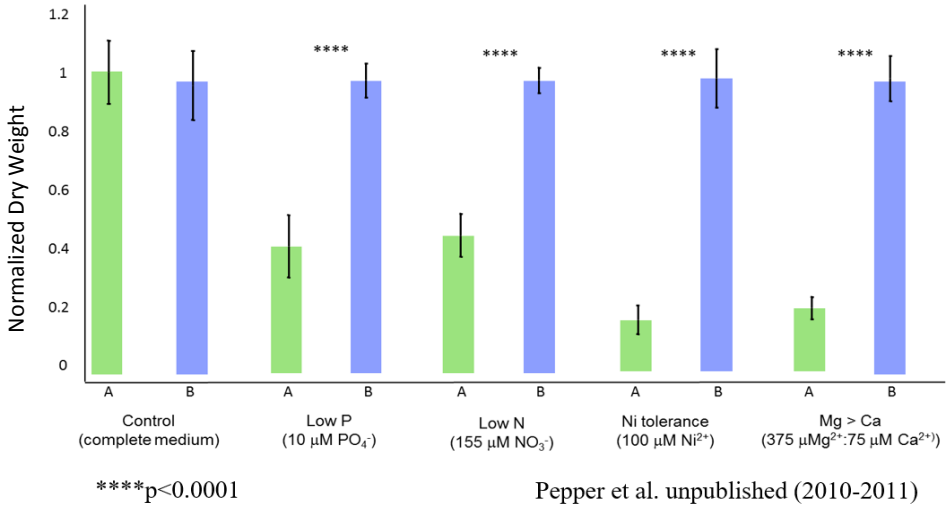

Figure 2. Proof-of-concept growth experiments using CAA and CAB. CAB (blue bars) outperformed (accumulated more dry mass) than CAA (green bars) in stressed conditions with $P < 0.0001$.

Serpentine soils can have phytotoxic levels of nickel that are often lethal to most plants (Yusuf et al., 2011) and can disrupt ecosystems (Chaney et al., 2008). Tolerance to high levels of

nickel is, without doubt, extremely important for some plants to be able to adapt to serpentine soils. One way serpentine plants deal with high levels of nickel is by hyperaccumulation (defined as accumulation of > 1000ug g-1 of dry weight); however, this is rare. Currently only 2% of serpentine species have met the requirements to be considered true hyperaccumulators (Kazakou et al., 2010). In laboratory growth experiments CAA and CAB were treated with varying amounts of nickel (0, 30, 60, 80, 100, 200, 500, or 1000 µmol/L $NiCl_2$). Again, CAB significantly outperformed CAA at concentrations of Ni > 30 µmol/L (Fig. 3). Representative tissues from each variation-treatment pair were analyzed using an energy-dispersive x-ray fluorescence (EDXRF) method and nickel concentrations were estimated.

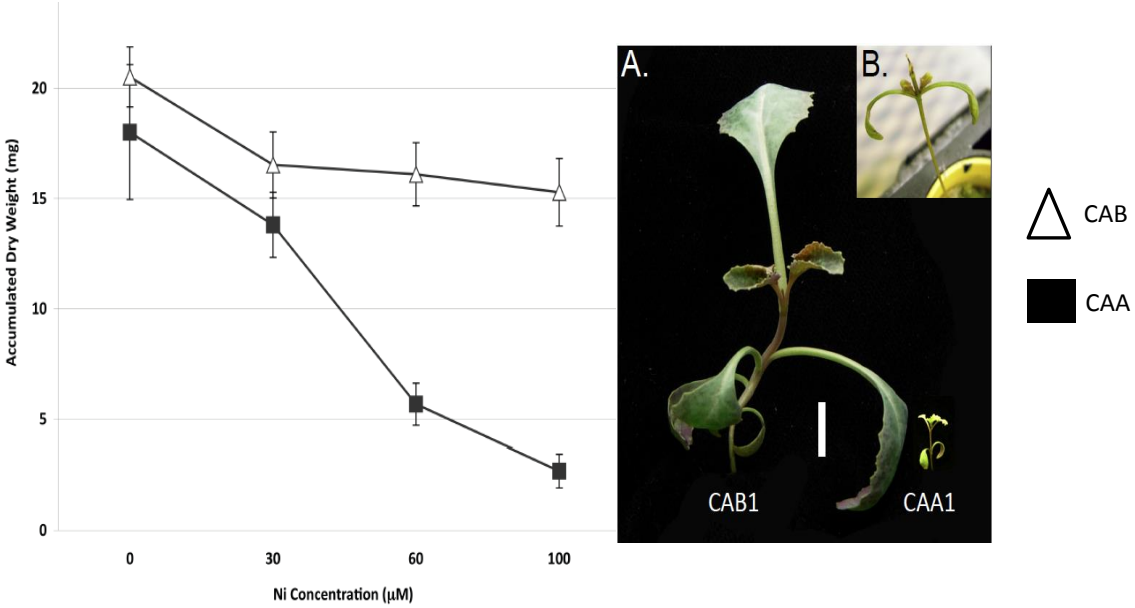

Figure 3. CAA and CAB grown in varying concentrations of Ni. A) A CAB plant at time of harvest (28 days) compared to B) A CAA plant of the same age at harvest date.

Hundreds of years of research have provided several examples of highly specialized plant adaptations to serpentine soils (Brooks, 1987); however, the molecular basis for these adaptations remain largely unknown. Advances in next generation sequencing technology and bioinformatics have allowed researchers to examine the genetic basis of adaptation using organisms that are traditionally considered 'non-model' species (Burrell et al., 2011). One major goal of the research in the Pepper lab is to generate a molecular arsenal of tools to warrant using CAB as the model organism for understanding the adaptation to serpentine soils. This project aims to elucidate the heritable differences in the CAA and CAB transcriptomes and genomes that may confer adaptation to serpentine soils.

Specific Aim 1: Characterization of the transcriptomes of CAA and CAB

Challenge: Transcriptome data was used to 1) detect single nucleotide polymorphisms (SNPs) and copy-number variants (CNVs) between CAA & CAB, 2) provide reference sequences for subsequent RNA-seq experiments (aim 2), 3) assist with the now complete functional annotation of the CAB whole genome sequencing in conjunction with the Joint Genome Institute (JGI), and 4) uncover global patterns of evolutionary trajectory using d*N*/d*S* measures of orthologs.

Approach: cDNA libraries were built for both variations grown in a variety of experimental conditions. Several bioinformatics tools were used, in conjunction, to perform *de novo* assembly of both transcriptomes. Transcripts were BLASTed to the non-redundant peptide (blastx) and nucleotide databases (blastn). Orthologs were uncovered using reciprocal best BLAST (RBH) approach and d*N*/d*S* was calculated on tentative orthologous pairs (TOPs). An all-by-all BLAST approach was implemented to detect paralogs within each taxa and d*S* values

between tentative paralogous pairs (TPPs) was used to look for patterns of complete or partial gene duplication events.

Impact: Transcriptome data is paramount for this project as they set the foundation for subsequent challenges (aim 2 and future research). The detection of SNPs and CNVs has provided novel genetic markers, and enabled predictions of functional differences between CAA and CAB gene products. Further, reference transcriptomes have led to quick and thorough analyses of a massive RNA-seq experiment (aim 2). Finally, transcriptome data has been used to functionally annotate the CAB1 and CAA1 whole genome sequencing projects.

Specific Aim 2: Global patterns of gene expression using RNA-seq technologies.

Challenge: Quantify expression level differences between CAA and CAB under varying environmental conditions (e.g., 1/4x MS media, granite soils, and serpentine soils). To detect allele specific expression, a F1 hybrid was also included in these experiments and will be analyzed in the near future when the CAA reference genome is more complete.

Approach: Seeds from each genotype (CAA, CAB, and a F1) were grown under all conditions mentioned above in three biological replicates. RNA was extracted from plants for subsequent use in library prep. Library sequencing was performed by the Hudson-Alpha Institute for Biotechnology using the Hi-Seq 2500 instrument. Gene expression levels were calculated as transcripts per million (TPM) by the program Salmon using the CAB1 genome assembly as reference and paired end fastq files of RNA-seq data from all replicates in all growth conditions. EdgeR was then used to calculate log fold change for all genes. Weighted correlation network analysis (WGCNA) was performed to cluster all genes based on their expression profile.

Impact: The RNA-seq data was separated into 18 modules (clusters) based on similar expression patterns. These modules were further processed using Fisher's exact test to find any overrepresented GO terms. Initial efforts were set out to answer the following three questions: 1) which genes are constitutively expressed in CAB, 2) which genes are induced in CAB on serpentine outcrops, and 3) which genes are induced in CAA on serpentine outcrops? RNA-seq data implicate a suite of chloroplast and plastid genes being constitutively expressed in CAB; this is an unexpected and novel finding. Genes induced in CAB on serpentine include those with roles in nutrient acquisition and movement and heavy metal binding, and genes induced in CAA on serpentine indicate response to nutrient starvation and galactose binding/transport. Ultimately results from these analyses, in conjunction with coding sequence data and QTL mapping, will be used to find high quality candidate genes that confer tolerance to serpentine soils.

Collaborative Efforts

1. Department of Energy Joint Genome Institute: Whole genome sequencing of *Caulanthus amplexicaulis* var. *barbarae* (CAB1) was performed. Both mate-pair and paired-end reads were assembled by Elyssa Garza (Pepper lab) using CLCbio and Allpaths software. Sequences were further annotated using the MAKER-P genome annotation pipeline. This method uses multiple data sources, in addition to genomic DNA, to build a near complete list of full length genes. Transcriptome data from aim 1 was used to help fully annotate the genome.

2. Sharon Strauss, U.C. Davis: Phenotypic analysis of recombinant inbred lines in native soils and on natural serpentine outcrops. The Strauss lab has performed growth experiments in native serpentine and granite soils as well as provided us with the soils to perform the RNA-seq analyses outlined in aim 2.

3. Undergraduate researchers, TAMU: Numerous undergraduate students have assisted and added to this dissertation. Two, in particular, Oscar Hernandez and Valerie Dietz provided novel data that was included in the transcriptome manuscript which warranted their inclusion as co-authors (see chapter 2). These students received hands-on molecular and bioinformatics training, and provided valuable data for chapter 2.

Broader Impacts

In addition to discovering the molecular basis for adaption to serpentine soils, this project and future projects may assist in a broader understanding of metal tolerance and accumulation and the efficient use of nitrogen and phosphorous in fertilizers.

1. Metal tolerance and accumulation: Phytoremediation, Greek for "plants restoring balance", is the removal of pollutants in soils, water, or air, with the aid of plants. Benefits of phytoremediation include traditionally lower costs than other remediation attempts, ease of plant monitoring, possible reuse of valuable metals via "phytomining" (Boonyapookana et al., 2005), and is often considered the least harmful method (over EDTA, etc.) because naturally occurring organisms are used and the environment can remain as close to 'natural' as possible. Phytoextraction, a subset of phytoremediation, removes environmental containments via hyper-accumulation and sequesters toxins in harvestable organs of plants that have evolved the capacity to not only survive, but sometimes thrive under these adverse conditions (Salt et al., 1998). A plant may be classified as a hyper-accumulator if it can absorb and sequester one of the following: Zn at 10,000 mg/kg, As, Ni, Cr, Pb at >1000 mg/kg, or Cd at >100 mg/kg (Wei et al, 2009). Hyper-accumulating plants can usually sequester only one specific metal (Gerhardt et al.,

2009; Hsiao et al., 2007); however serpentine plants have adapted to a suite of heavy metals, and are thus ideal organisms to examine metal tolerance and accumulation. Burrell et al. (2012) showed that CAB hyper-accumulate nickel to extraordinary levels (~1% dry weight). Understanding the molecular mechanisms influencing metal tolerance/accumulation could lead to engineering larger plants for bigger and better phytoremediation attempts.

2. Efficient use of nitrogen and phosphorous in plant fertilizers: As the human population continues to increase, so do efforts to provide enough food. Over the past 50 years, the production of crop plants has doubled, leading to a 7-fold increase in nitrogen levels from fertilizers (Hirel et al., 2007). Most of the worlds' phosphorous is mined from phosphate rock, a non-renewable resource, and as levels are decreasing, associated mining costs are increasing (Cordell et al., 2008). Phosphate mining also produces heavy metal pollution. However, nitrogen and phosphorous are essential plant nutrients; most crop species use these nutrients very inefficiently which necessitates the use of fertilizers in very large quantities. Much of the N and P applied to crop plants are not used by the plants and result in run off in surrounding bodies of water, leading to eutrophication of both freshwater and marine systems (Hirel et al., 2007). CAB plants show significantly better growth compared to CAA in experiments limiting or completely removing nitrogen and phosphorous (Fig. 2). Understanding how CAB plants can tolerate such low levels of essential nutrients, and elucidating the genes responsible, will provide unprecedented insights into engineering more efficient crop plants that will require a fraction of the fertilizers they now receive.

Perspectives

In all, this dissertation has two major experiments: 1) the assembly and annotations of comprehensive reference transcriptomes for both CAA and CAB and 2) RNA-seq analyses of CAA and CAB to reveal gene expression patterns in various environmental conditions. The transcriptome experiment is complete and was published in Genome Biology and Evolution (GBE) in December, 2017. The RNA-seq experiment was a massive undertaking and resulted in even more data. This dissertation addresses three major comparisons to hone in on and answer what is believed to be the most important questions regarding gene expression among different genotypes and on varying soils.

CHAPTER II

TRANSCRIPTOME SIGNATURES OF SELECTION, DRIFT, INTROGRESSION, AND GENE DUPLICATION IN THE EVOLUTION OF AN EXTREMOPHILE PLANT*

Overview

Plants on serpentine soils provide extreme examples of adaptation to environment, and thus offer excellent models for the study of evolution at the molecular and genomic level. Serpentine outcrops are derived from ultramafic rock and have extremely low levels of essential plant nutrients (e.g., N, P, K, Ca), as well as toxic levels of heavy metals (e.g., Ni, Cr, Co) and low moisture availability. These outcrops provide habitat to a number of endemic plant species, including the annual mustard *Caulanthus amplexicaulis* var. *barbarae* (*Cab*) (Brassicaceae). Its sister taxon, *C. amplexicaulis* var. *amplexicaulis* (*Caa*), is intolerant to serpentine soils. Here, we assembled and annotated comprehensive reference transcriptomes of both *Caa* and *Cab* for use in protein coding sequence comparisons. A set of 29,443 reciprocal best Blast hit (RBH) orthologs between *Caa* and *Cab* was compared to identify coding sequence variants, revealing a high genome-wide d*N*/d*S* ratio between the two taxa (mean = 0.346). We show that elevated d*N*/d*S* likely results from the composite effects of genetic drift, positive selection, and the relaxation of negative selection. Further, analysis of paralogs within each taxon revealed the signature of a period of elevated gene duplication (~10 Ma) that is shared with other species of the tribe Thelypodieae, and may have played a role in the striking morphological and ecological

---

*Reprinted with permission from Transcriptome signatures of selection, drift, introgression, and gene duplication in the evolution of an extremophile endemic plant by Angela K. Hawkins, Elyssa R. Garza, Valerie A. Dietz, Oscar J. Hernandez, W. Daryl Hawkins, A. Millie Burrell, and Alan E. Pepper. 2017. *Genome Biology and Evolution* Volume 9, Issue 12.
https://doi.org/10.1093/gbe/evx259

diversity of this tribe. In addition, distribution of the synonymous substitution rate, d*S*, is strongly bimodal, indicating a history of reticulate evolution that may have contributed to serpentine adaptation.

Key words: adaptation, *Caulanthus*, serpentine, *Streptanthus*, Thelypodieae, ultramafic

Introduction

As sessile organisms, plants provide excellent models for both field-based and laboratory studies of fine-scale adaptation to environment. Outcrops of serpentine geology are one of the most extreme environments encountered by land plants. These habitats are low in essential mineral nutrients (e.g., N, P, K, Ca) and have high levels of toxic heavy metals (e.g., Ni, Cr, and Co). Serpentine-derived soils are shallow, poorly developed, lack organic matter, and are prone to moisture limitation (Whittaker 1954b; Brady, et al. 2005; Kazakou, et al. 2008). Plants in these environments are also subject to high-light and elevated-temperature stresses due to sparse vegetation and low community-level evapotranspiration. A small number of plants that have adapted to outcrops of serpentine geology provide compelling examples of natural selection in response to complex environmental challenges. However, the genetic, molecular, and physiological mechanisms underpinning these remarkable adaptations are largely unknown.

Several naturally occurring examples of conspecifics with extreme differences in habitat preference (e.g., serpentine tolerant vs. intolerant) allow highly informative reciprocal transplant experiments (Kruckeberg 1951). In addition, recent applications of population genomics to natural serpentine and non-serpentine plant populations show significant promise as a tool for uncovering the genetic mechanisms of serpentine tolerance (Turner, et al. 2010; Arnold, et al. 2016). Further, some serpentine-related phenotypes, such as tolerance to nickel, can be studied

in laboratory settings and are amenable to intensive genetic analyses in controlled environments (Burrell, et al. 2012).

Serpentine soils have poor biological productivity resulting in reduced vegetation density and high rates of edaphic endemism (Whittaker 1954b; Kruckeberg and Rabinowitz 1985; Safford, et al. 2005). For example, roughly 1.5% (6,000 km^2) of the California Floristic Provence is serpentine, yet these areas support ~13% of all endemics in this flora (Pepper and Norwood 2001). As with serpentine tolerance, the underlying causes of this pattern of endemism are unknown. *Caulanthus amplexicaulis* var. *barbarae* (J. Howell) Munz (designated *Cab*), is an annual diploid plant that is entirely restricted to a series of isolated serpentine outcrops in the San Rafael Mountains of coastal California (Howell 1962; Safford, et al. 2005). Its sister taxon, *C. amplexicaulis* var. *amplexicaulis* S. Watson (designated *Caa*), is intolerant to serpentine soils and is found mainly on open granite outcrops throughout the Transverse Ranges of southern California. The two varieties can readily hybridize in controlled crosses (Burrell, Taylor, et al. 2011), but in nature they are geographically isolated, with the closest populations separated by ~75km. Phylogenetic analyses (Pepper and Norwood 2001) support a biotype-depletion evolutionary model (Stebbins 1942) in which *Caa* and *Cab* are descended from a more generalist ancestor that may have transitioned to granite and serpentine outcrops as refugia from competition (Anacker 2014).

Since *Cab* is strictly endemic to serpentine, we anticipated that populations would be largely fixed at those loci that are most critical for adaptation to serpentine. At the molecular level, key allelic differences between *Cab* and *Caa* might include variants in transcriptional regulatory sequences (*cis*-acting sites), splicing, gene copy number, and protein coding sequences. Here, we employed deep sequencing to obtain comprehensive reference

transcriptomes for representatives of *Caa* and *Cab* (designated CAA1 and CAB1, respectively (Pepper and Norwood 2001; Burrell, Taylor, et al. 2011) to use to compare protein coding genes and provide a reference for comparative analyses of transcript abundance and mRNA structure.

A key objective of this work was to use patterns of coding sequence evolution to aid in the identification of loci evolving under positive selection. However, evolutionary outcomes at the molecular level are determined by a diverse set of mechanisms that includes both natural selection and non-selective processes such as hybridization, genetic bottlenecks, founder effects, and genetic drift due to small population size (N_e). As a rare endemic species, *Cab* typically occurs in geographically isolated clusters with <100 individuals at reproductive maturity (with some clusters having <20 individuals over multiple years). Based on microsatellite markers, estimates of N_e for *Cab* are in the single- and low double-digits, and gene flow among *Cab* populations is limited (Burrell et al., unpublished). *Caa* is endemic to the Transverse Ranges of southern California and is largely restricted to open, newly-eroded granite-derived talus slopes. The two populations of *Caa* that have been examined in similar detail (Burrell et al., unpublished) also show small N and N_e values. Thus, any model for the evolutionary history of the two taxa must consider and reconcile the effects of both rigorous natural selection and small population sizes.

In this work, we compared orthologs of *Caa* and *Cab*, and found an unexpectedly high d*N*/d*S* ratios across much of the genome. We explored several possible explanations for this phenomenon, including positive selection. Further, this work revealed unexpected complexity in the evolutionary pathways leading to the divergence of the two taxa, including signatures of recent gene duplication and introgression – processes that may have contributed to adaptive evolution.

Materials and Methods

Plant materials and growth conditions

This study utilized inbred lines that were representative of *Caulanthus amplexicaulis* var. *amplexicaulis* (CAA1) and *Caulanthus amplexicaulis* var. *barbarae* (CAB1) (Pepper and Norwood 2001; Burrell, No, et al. 2011; Burrell, Taylor, et al. 2011). Because of the combination of a strong selective regime along with small population sizes, we made the assumption that the alleles most critical for serpentine adaptation would be fixed, and thus present in the inbred exemplar lines. Further, we anticipated that the use of highly homozygous inbred lines would greatly simplify *de novo* assembly by reducing the difficulties in distinguishing alleles from paralogs in sequencing data. The CAA1 line was obtained through selfing of seeds collected on a granite outcrop in Los Angeles County, CA, USA, while the CAB1 line was obtained from a serpentine barren in Santa Barbara County, CA, USA (Pepper and Norwood 2001). These source locations were matched as closely as possible with regard to elevation, latitude, annual precipitation, and slope/aspect. Thus, the key environmental differences were presumed to be the physical and chemical properties of the source soils.

To obtain tissues for RNA isolation, both taxa were grown in growth chambers in a variety of environmental conditions, and several organs and tissues were harvested from each taxon (supplementary table S1, Supplementary Materials Online). As base media for manipulation of nutrient conditions, we used Murashige and Skoog (MS) medium, ¼ strength, with salts, micro-, and macronutrients, pH 5.8 (Burrell, et al. 2012). Floral and fruit tissues were not obtained from CAB1 because of its later flowering time in the laboratory conditions employed.

The *A. thaliana* confirmed homozygous *phl1* mutant line SALK_079505C was obtained from the Salk collection of indexed T-DNA insertion lines (Alonso, et al. 2003). To test for growth in limiting phosphate, CAA1, CAB1, wild-type *Ath* Col-0, and SALK_079505C were grown in 1/4x MS media, as described, but with varying concentrations of KPO_4 for 28 d (*Caulanthus*) or 22 d (*Ath*) after the emergence of first true leaves. Whole aerial portions of plants were harvested and dry biomass was measured as a proxy for fitness.

RNA isolation and transcriptome sequencing

Upon harvest, tissues were immediately stored in liquid N_2. RNAs were extracted using the RNAqueous® total RNA isolation kit (ThermoFisher). Genomic DNA was removed using the TURBO DNA-free kit (ThermoFisher). RNA integrity was assessed using the Agilent Bioanalyzer (Agilent Technologies, Santa Clara, CA); all samples had a RIN (RNA Integrity Number) score of 7.0 or greater. The SMARTer® PCR cDNA Synthesis kit (Clontech, Mountain View, CA) was used to reverse transcribe total RNA into cDNA using a polyA-specific primer. A duplex-specific nuclease (Evrogen, Moscow, Russia) was used to normalize cDNA libraries (Zhulidov, et al. 2004) which were sent to the Genomic Sequencing and Analysis Facility (GSAF) at the University of Texas, Austin for paired-end (2x100 bp) sequencing using the Illumina Hi-Seq 2000 instrument. Sequencing reads were submitted to the NCBI short read archive under the BioProjects PRJNA417948 (*CAA1*) and PRJNA417949 (*CAB1*).

De novo transcriptome assembly

Raw IIlumina reads were processed using Trimmomatic (Bolger, et al. 2014) to trim for quality (Q20) and a minimum length of 50 bp. To mitigate the potential effects of sequencing errors

introduced by PCR, we removed duplicate reads using the CLCBio Genomics Workbench (CLCBio, Cambridge, MA, USA). It has been previously shown that the Trinity *de novo* assembly pipeline (Grabherr, et al. 2011) and the Velvet/Oases pipeline (Zerbino and Birney 2008; Schulz, et al. 2012) can each produce unique (i.e., non-overlapping) transcripts from the same set of sequencing reads (Devisetty, et al. 2014). For this reason, we assembled reads independently using both pipelines and then merged the resulting transcript sets. Specifically, reads were assembled using Velvet v1.2.10 followed by Oases v0.2.08. Independent runs of Velvet were performed at k-mer values of 21, 27, 31, 37, 41, 47, 51, 57, 61, and 67, as it has been shown that individual genes require different k-mer and coverage cut off values to be assembled correctly (Gruenheit, et al. 2012). Assemblies were merged using Oases at a k-mer value of 61. A Python script was used to select the most reliable transcript per locus at a cutoff fraction of 0.9 (Reich https://code.google.com/archive/p/oases-to-csv/). Reads were independently assembled using Trinity r20140717 with default parameters (e.g., k-mer fixed at 25). CD-HIT-EST (Li and Godzik 2006) was used to merge the Velvet and Trinity assemblies using a 0.95 similarity threshold and word size of 10 (all other parameters were set to default), thus removing redundant transcripts and yielding a set of high-confidence representative transcribed loci (RTL) for each taxon. Plastid-derived transcripts were compared to long-read genomic sequences (Burrell, No, et al 2011) to correct for RNA editing. Assembled representative transcribed loci (RTL) were submitted to the NCBI Transcriptome Shotgun Assembly (TSA) database under the projects GGBY00000000 (CAB) and GGBZ00000000 (CAA).

Functional annotation

RTL were queried against the NCBI non-redundant protein (nr) and nucleotide (nt) databases using e-value thresholds of 10^{-8} for BLASTN and 10^{-6} for BLASTX. RTL were also queried against the TAIR10 CDS database using BLASTN and BLASTX (at the same thresholds) to find the best hit to the *Arabidopsis thaliana* genome. RTL were processed using Blast2Go v3.1.3 (Conesa, et al. 2005) to assign gene ontology (GO) terms using a BLASTX search of the NCBI nr subset Viridiplantae (taxid: 33090), with a word size of 3, HSP cuttoff of 33, and e-value $<10^{-10}$. GO enrichment analyses were performed using Fisher's Exact Test. In this test, any positive results from GO terms with less than five loci in either the test or reference set were ignored. Enrichment results were processed through GO Trimming version 2.0 (Jantzen, et al. 2011) to reduce redundancy in GO terms.

Identification of orthologous loci

The reciprocal best BLAST-hit (RBH) method has been found to out-perform a number of orthology identification algorithms (Altenhoff and Dessimoz 2009). To identify orthologs between CAA1 and CAB1 transcripts, we performed RBH analysis of RTL using Perl scripts developed by the Systems Biology Research and Resources group at Harvard University. Tentative orthologous pairs (TOPs) between CAA1 and CAB1 were analyzed using the TRAPID pipeline (Van Bel, et al. 2013) to identify the longest ORF for each transcript. RBH was also used to identify the best putative ortholog for each TOP in the *Arabidopsis thaliana* TAIR10 CDS database.

Analyses of coding sequence evolution

To detect variants between TOPs, CAA1 reads were mapped to the reference CAB1 transcripts using the 'map reads to reference' function of the CLC Genomics Workbench 7.02, with a mismatch cost of 2, indel cost of 3, minimum length fraction of 0.5, and minimum similarity fraction of 0.8. The 'basic variant detection' algorithm of the CLC Genomics Workbench 7.0.2 was implemented using a ‘haploid’ ploidy model with minimum coverage of 10, minimum count of 10, and minimum frequency set to 90%. To estimate pairwise synonymous and non-synonymous substitution rates in TOPs, we used the yn00 maximum likelihood utility from the PAML package (Yang 1997), implementing the counting method (Yang and Nielsen 2000).

For GO enrichment analyses, TOPs were separated into bins of d*N*/d*S* ranges 0–0.099, 0.1–0.299, 0.3–0.799, 0.8–1.199, and >1.2. The bin with d*N*/d*S* of >1.2 was used to identify genes under possible selection. The category between 0.8 and 1.199 presumably included genes evolving under neutrality (~1.0), and would also include genes in which some residues are under positive selection while others are constrained by negative selection. The remaining categories were partitioned so that each bin included a similar number of genes in order to provide a similar level of statistical power to detect enrichment.

For genes in which d*N*/d*S* ratio could not be calculated because d*S* = 0, a novel 'synthetic' d*N*/d*S* metric designated ω_S was calculated as the ratio of the observed d*N* divided by a d*S* value that would be obtained if there were a single synonymous nucleotide substitution in the alignment (i.e., d*S* > 0). For this analysis, we considered a threshold of >1.2 to be analogous to a d*N*/d*S* ratio of >1.2 as a heuristic for detection of genes possibly evolving under positive selection. For GO enrichment analyses of this gene set, TOPs were separated into two categories,

a test file including TOPS with ω_s >1.2 and a reference file with ω_s values under 1.2, and analyzed with a FDR threshold of 0.05.

Identification of parologous loci

Transcripts from both CAA1 and CAB1 were subjected to all-against-all BLASTN searches (CAA1 vs. CAA1; CAB1 vs. CAB1) with a threshold e-value $<10^{-10}$, at least 60% identity, and ORFs of >300 bp (Blanc and Wolf, 2004). Within each taxon, the yn00 method from PAML (Yang 1997) was used to calculate d*N*, d*S*, and d*N*/d*S* ratios between paralogs.

Estimation of time of divergence

Bayesian estimation, implemented by BEAST v1.8.3 (Drummond, et al. 2012), was used to estimate time of divergence between *Caa* and *Cab*. Within any one lineage, substitution rates vary greatly between nuclear, mitochondrial, and plastid genomic compartments (Drouin, et al. 2008) and relative rates within the each compartment can vary dramatically among different evolutionary lineages (Smith and Klicka 2013; Hertweck, et al. 2015). To make meaningful comparisons of divergence time estimates from different compartments, we employed independent BEAST analyses using orthologous plastid and nuclear genes of *A. thaliana* as the outgroup sequences. Concatenated sets of 56 plastid and 24 randomly-selected nuclear genes were used to obtain divergence time estimates for their respective genomic compartments. The general time reversible (GTR) substitution model was implemented along with a strict molecular clock. The MCMC burn-in was set to 100 million, and parameters were resampled every 10,000 generations.

Molecular dating of plants in general, and the Brassicaceae family in particular, suffers from a paucity of reliable fossils that can be used for calibration. For this reason, the age of the Brassicaceae remains uncertain, with estimates that vary from ~54 Ma (Beilstein, et al. 2010) to ~32 Ma (Hohmann, et al. 2015). In this study, we employed a framework based on more recent dates of origin and divergence in the Brassicaceae (Franzke, et al. 2016) and used an estimated time since divergence of Brassicacae Lineage I (e.g., *Arabidopsis*) from Lineage II (e.g., *Caulanthus*) of 23.4 ± 0.7 Ma (Hohmann, et al. 2015) as a single calibration point.

Results

Assembly and annotation of reference transcriptomes

CAA1 and CAB1 RNAs were isolated from a variety of tissues, at various stages of plant development, and under differing environmental conditions (supplemental table 1, in Appendix A). RNA samples were pooled for each taxon and then used to create a pair of comprehensive normalized cDNA libraries for deep sequencing, yielding ~80 million filter-passed IIlumina paired-end reads for CAA1 and ~50 million paired-end reads for CAB1. Trimmed reads were assembled *de novo* and merged to yield a non-redundant set of 93,647 representative transcribed loci (RTL) for CAA1 (N50 = 1,001 bp), and 83,484 RTL for CAB1 (N50 = 691 bp) (supplemental table 2, in Appendix A). Approximately 68% of the RTL from both transcriptomes had significant BLASTN hits (*e*-value $<10^{-8}$) to sequences in the NCBI non-redundant nucleotide database (nt). Similarly, 74% of CAA1 RTL and 80% of CAB1 RTL had BLASTX hits at an *e*-value $<10^{-6}$ in the NCBI non-redundant protein database (nr). Of those RTL that had significant BLASTN hits to the nr database, 99% of CAA1 and 98% of CAB1 RTL had top hits to plant species; among these, 98% and 97%, respectively, had top hits to members

of the Brassicaceae family. Both CAA1 and CAB1 had BLASTX hits to 100% of the set of 248 CEGMA core eukaryote genes (Parra, et al. 2007) at a threshold *e*-value of $<10^{-20}$, indicating a high level of representation of expressed genes in both libraries.

Coding Sequence Evolution in Orthologous Gene Pairs

Reciprocal Best Blast-Hit (RBH) analysis of RTLs produced 29,443 tentative orthologous pairs (TOPs) between CAA1 and CAB1 with a threshold *e*-value $< 10^{-20}$ and a minimum alignment length of 150 bp. Of these, 84% of sequence pairs had significant BLASTN hits to the NCBI nt database (*e*-value $<10^{-8}$) and 87% had BLASTX hits to the NCBI nr database (*e*-value $< 10e^{-6}$). For variant calling, CAA1 reads were mapped to the CAB1-derived merged assembly, yielding 53,359 SNPs (frequency = 0.045%) and 848 indel polymorphisms (242 of these produced putative frameshifts; supplemental table 3, in Appendix A). In addition, 284 TOPs (~1%) were identified as having SNP polymorphisms involving loss or gain of a stop codon (e.g., nonsense polymorphisms). A further set of 4,108 TOPs (13.4%) had identical sequences between CAA1 and CAB1 (i.e. $dN = dS = 0$).

Pairwise PAML analysis between CAA1 and CAB1 of TOPs with ungapped alignments yielded a genome-wide mean dN/dS ratio of 0.346 (median = 0.220; Fig. 4). This value was relatively high compared to other species pairs across a wide range of taxonomic groups (table 1). Further, an unexpectedly large fraction of the TOPs (1,041 or ~3.5%) had dN/dS ratios of > 1.2 —an arbitrary threshold value that we presumed would be enriched in loci under positive selection. Of these, 835 TOPs (77.5%) had significant hits to a reference database (nt, nr, or TAIR10 CDS; supplemental table 4, attached file). GO annotation assigned these loci to a

number of functional categories, including several with possible ecological roles in serpentine tolerance, such as 'ion binding', and 'transmembrane transporter' (fig. 5).

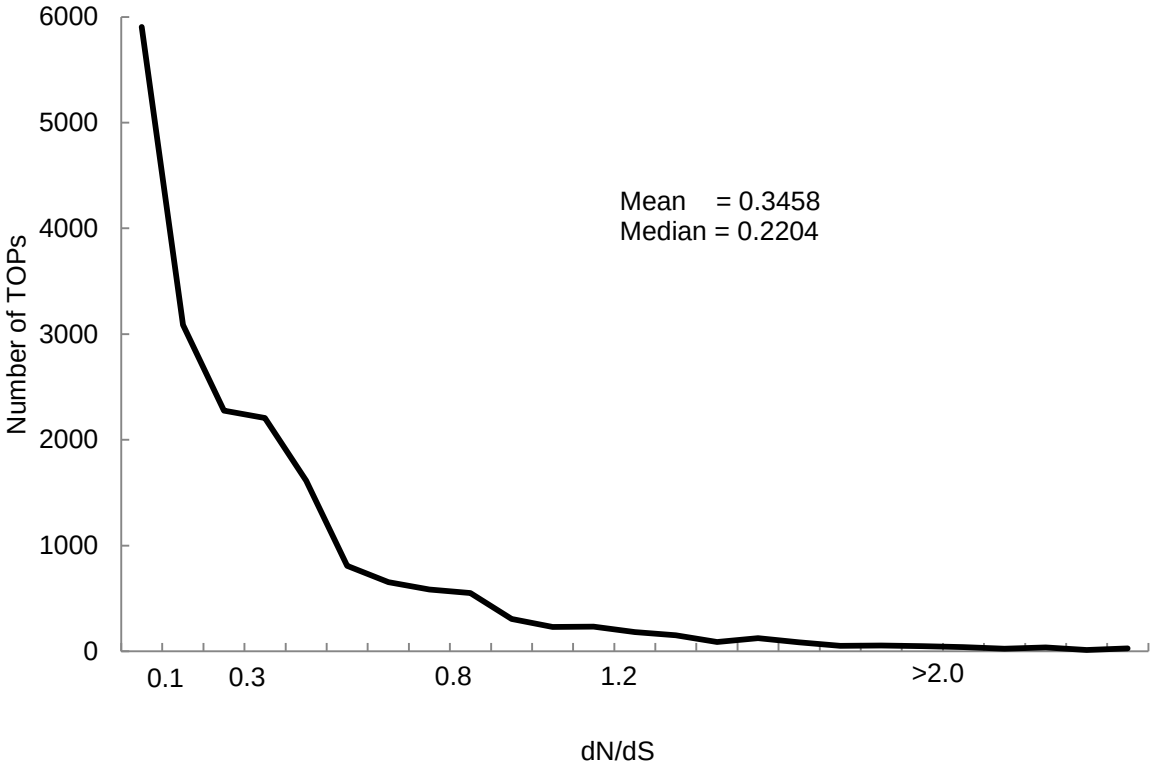

Figure 4. Histogram of d*N*/d*S* ratios of TOPs from CAA1 and CAB1. Values were calculated using the yn00 method in PAML.

In addition, there were 3,144 TOPs that had a d*N* > 0, but the d*N*/d*S* ratio was undefined because d*S* = 0. In most studies examining d*N*/d*S*, such genes are ignored. Given the relatively

Table 1. List of mean d*N*/d*S* ratios from available datasets. A representative sampling of global mean d*N*/d*S* ratios from EST, transcriptome, and genome sequencing datasets.

Taxa compared	Genes	d*N*/d*S*	Citation
White Oak spp.	28,676	0.32-0.38	(Cokus, et al. 2015)
Caulanthus amplexicaulis	29,433	0.35	This study
Whitefly sspp.	3,585	0.23	(Wang, et al. 2011)
Human/Chimp	13,198	0.20	(Arbiza, et al. 2006)
Yak and Cattle	8,923	0.18	(Qiu, et al. 2012)
Cichlid fish spp.	13,106	0.17	(Elmer, et al. 2010)
Arabidopsis/Brassica	310	0.14	(Tiffin and Hahn 2002)
Cephalochordate spp.	8,333	0.12	(Yue, et al. 2014)
Pufferfish spp.	16,950	0.11	(Montoya-Burgos 2011)
Teleost Fish spp.	4,033	0.10	(Ren, et al. 2014)

recent divergence of CAA1 and CAB1 (Pepper and Norwood 2001), and thus low expected *dS* values, we surmised that this category might include loci with non-synonymous changes of adaptive significance. To identify loci under possible positive selection from within this category, we developed a novel 'synthetic' d*N*/d*S* metric ω_s that was based on the premise that the first (hypothetical) synonymous substitution to occur would result in a calculable d*N*/d*S* ratio. Thus, for each locus we introduced a single synonymous mutation *in silico* to give a positive value for d*S* and then recalculated d*N*/d*S*. By this metric, 549 of these 3,144 TOPs (17.5%) had ω_s values >1.2 (supplemental table 4B, attached file).

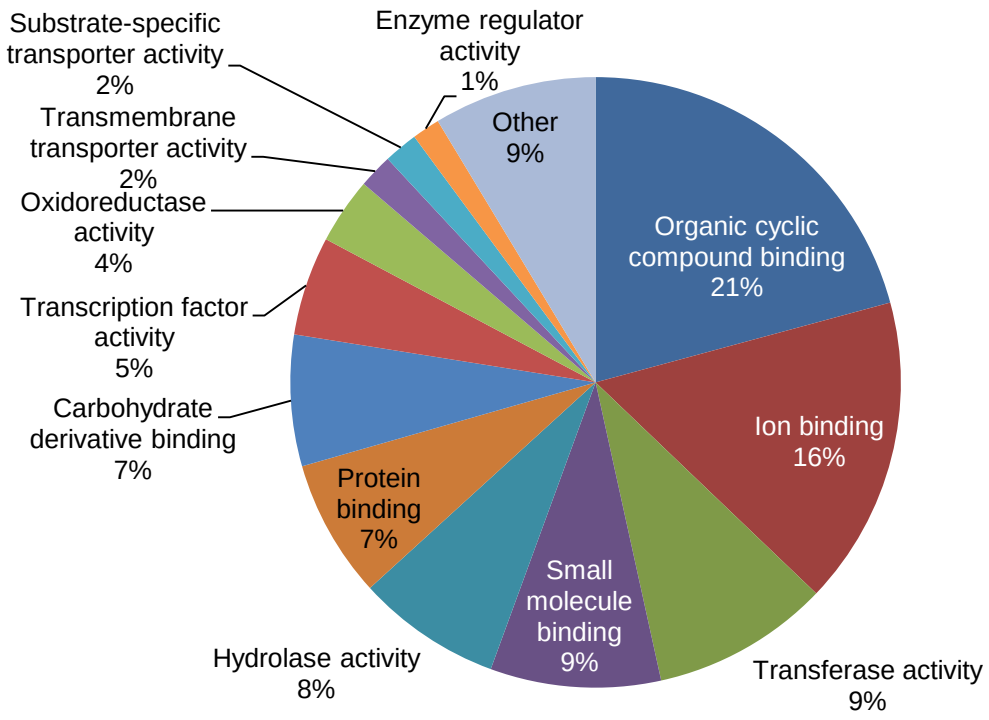

Figure 5. Distribution of level 3 GO terms for TOPs with high d*N*/d*S*. Terms for molecular function among TOPs with d*N*/d*S* > 1.2.

Effects of population size

Although the finding of protein encoding genes with high d*N*/d*S* ratios is widely attributed to positive selection (Nielsen 2005), there are a number of viable alternative explanations. Classical genetic theory and recent empirical studies have demonstrated that genetic drift reduces the efficacy of negative selection in the purging of weakly deleterious alleles, resulting in elevated d*N*/d*S* ratios (Wright 1931a; Eyre-Walker, et al. 2002; Strasburg, et al. 2011). To explore this effect, we compared pairwise d*N*/d*S* values of CAA1 and CAB1, which have population sizes (N and N_e) of less than 10^2, with those of *Arabidopsis lyrata* and *Capsella grandiflora* (Brassicaceae), which diverged from each other ~8 Ma (Hohmann, et al. 2015) and have much

larger populations sizes. *C. grandiflora* is an obligate outcrossing plant noted for large effective population sizes (N_e ~10^5–10^6) and very little population structure (Gossmann, et al. 2010; St Onge, et al. 2011). Similarly, *A. lyrata* is also an obligate outcrosser that has estimated N_e values in the range of ~10^3–10^4 (Ross-Ibarra, et al. 2008).

For this comparison, we employed a set of 218 orthologous loci that had been previously selected without regard to gene function (Ross-Ibarra, et al. 2008). This gene set was employed because the effects of positive and negative selection have been extensively studied in both *C. grandiflora* and *A. lyrata*, and the biological functions of these genes are largely known (Ross-Ibarra, et al. 2008; Slotte, et al. 2010). RBH was used to identify 177 sets of 1:1:1:1 orthogroups with ungapped alignments from the four taxa. Within this set of genes, the mean d*N*/d*S* ratio for the CAA1 vs. CAB1 comparison (0.238) was significantly higher (P = 0.001 in a Wilcoxon paired signed-rank test) than that of *A. lyrata* vs. *C. grandiflora* (0.127). The median values differed similarly (0.136 for CAA1 vs. CAB1, 0.069 for *A. lyrata* vs. *C. grandiflora*, P = 0.0001), Further, the elevated d*N*/d*S* values in CAA1 vs. CAB1 occurred across a broad distribution of genes rather than in just the highest d*N*/d*S* categories (fig. 6). Taken together, these results indicate that the observed difference in mean d*N*/d*S* in the two comparisons was due to a large proportion of the ortholog pairs having higher d*N*/d*S* values in CAA1 vs. CAB1 (rather than a few outliers with high d*N*/d*S*, as would be expected, for example, in cases of misalignment). Since the gene set used in this comparison was selected arbitrarily, it is unlikely that this pattern of broadly elevated d*N*/d*S* arose entirely from either widespread positive selection or relaxation of selection. Rather, this pattern was more consistent with broadly reduced efficacy of purifying selection due to genetic drift, which would be expected to affect a wide range of functional categories.

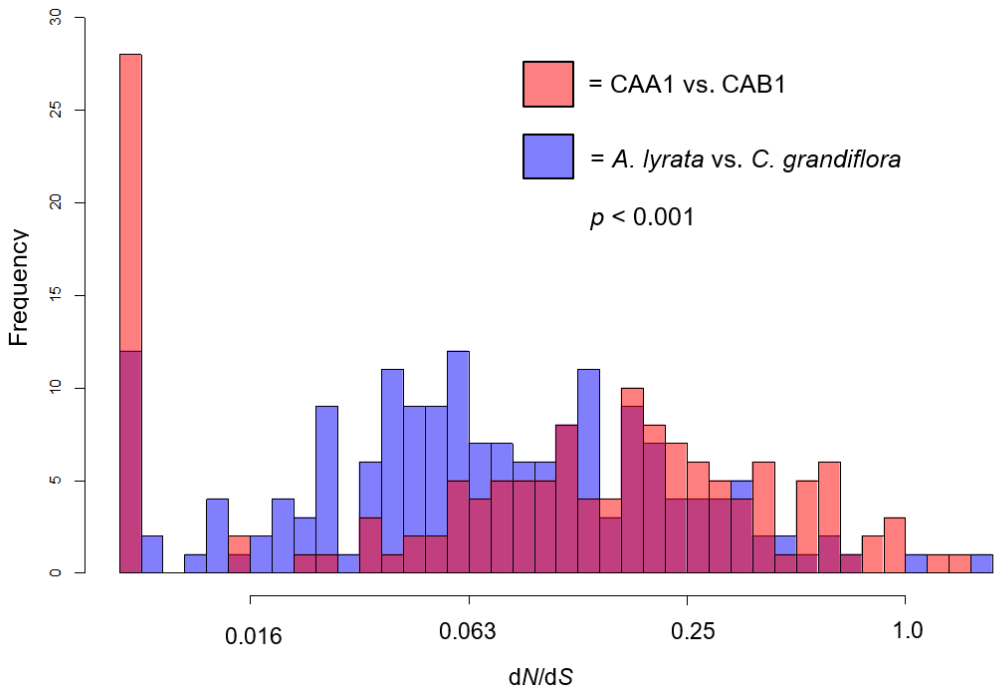

Figure 6. Histogram comparing d*N*/d*S* ratios between two sets of taxa. Overlay histogram comparing d*N*/d*S* ratios between CAA1 versus CAB1 (red) and *Capsella grandiflora* versus *Arabidopsis lyrata* (blue) based on 177 1:1:1:1 orthogroups. Column at far left corresponds to orthologs with a d*S* value of 0. Pink corresponds to overlap between taxa sets.

Evidence for relaxation of negative selection

Relaxation of negative selection occurs in certain ecologically relevant genes when an organism colonizes a novel environment. In this scenario, some genes may become dispensable (Lahti, et al. 2009) and show a trend toward neutrality (i.e. d*N*/d*S* ~1). Both *Caa* and *Cab* occur on rocky barrens in sparsely distributed populations, with little or no intra- or interspecific competition for light. One would expect that in both *Caa* and *Cab*, genes in the red/far-red shade avoidance pathway, which confers adaptive phenotypic plasticity of growth form and flowering in responses to the proximity of competitors (Casal 2013), might show evidence of weakened purifying selection. To test for possible signatures of relaxation of negative selection resulting

from the transition from more general habitats to the specialized granite and serpentine environments, we examined the coding sequence evolution of the 20 loci in the 'shade-avoidance' GO category GO:0009641 (table 2).

Table 2. d*N*/d*S* values of CAA1 and CAB1 genes in the shade avoidance GO category.

Ath ortholog	TOP	*e*-value	d*N*/d*S*	*Ath* gene annotation
AT1G06040	—	NA	NA	*STO/BBX24*
AT1G10390	+	0.00	0.304	*DRA2*
AT1G18400	+	2.62E-168	1.089	*BEE1*
AT1G70560	+	3.88E-165	0.366	*TAA1*
AT1G73870	—	NA	NA	*BBX16/COL7*
AT1G75540	+	2.92E-10	0.500	*BBX21*
AT1G78600	+	0.00	0.355	*LZF1*
AT1G80360	+	1.40E-64	0.214	*ISS1/VAS1*
AT2G32950	+	0.00	0.000	*COP1*
AT2G39940	+	0.34E-49	0.000	*COI1*
AT2G42870	+	2.09E-63	1.387	*PAR1*
AT2G44910	—	NA	NA	*ATHB4*
AT2G46970	—	NA	NA	*PIF3-like 1*/*PIL1*
AT3G58850	+	2.82E-122	0.422	*PAR2*
AT4G08920	+	0.00	0.000	*CRY1*
AT4G16780	+	0.00	0.178	*ATHB2*
AT4G25260	+	0.00	0.000	*Pectinesterase inhibitor*
AT4G31500	+	4.41E-85	0.400^{s}	*SUR2/RNT1/RED1*
AT5G08130	+	7.99E-89	0.456	*BIM1*
AT5G47370	+	0.00	0.124	*HAT2*

Note—*Ath* ortholog refers to the *Arabidopsis thaliana* gene models from GO:0009641 category; TOP indicates the presence or absence of an orthologous TOP from CAA1/CAB1, established by RBH to *Ath* CDS; e-value is obtained from BLASTN search of TOP consensus sequence to *Ath* CDS; d*N*/d*S* values are between CAA1 and CAB1 orthologs; annotation refers to mutant or gene names from TAIR. ^{s} Refers to synthetic dN/dS ratio (ω_s).

Within this gene set, TOPs for four genes could not be identified. For one of these genes, *BBX16* (At1G73870), a *CONSTANS*-like zinc finger transcription factor, no orthologous transcripts were assembled from either CAA1 or CAB1. Based on BLAST searches of available genome databases, orthologs of this *Ath* gene are apparently absent from all of Lineage II of the Brassicaceae family (not shown). The 16 remaining loci had calculable d*N*/d*S* ratios with a mean of 0.356 (median = 0.304).

Four of these genes were apparently evolving under strong purifying selection (d*N*/d*S* = 0.0). These included: 1) the centrally important photoreceptor *CRY1* (At4G08920), which has numerous roles beyond shade avoidance, including blue-UVA stimulation of stomatal opening and phototropism (Chaves, et al. 2011), 2) jasmonate receptor *COI1* (At2G39940), which plays a critical role in defense responses (Xie, et al. 1998), and 3) *COP1* (At2G32950), which when mutated shows a severely pleiotropic dwarf phenotype (Kwok, et al. 1996). The fourth locus encoded a putative pectinesterase inhibitor (At4G25260) that has postulated roles in defense, cold-acclimation, and hormone responses (Goda, et al. 2004; Brodersen, et al. 2006; Oono, et al. 2006) as well as shade avoidance. The removal of these four pleiotropically-acting genes, the mean d*N*/d*S* value increased to 0.470 (median = 0.360), with two genes evolving under apparent neutrality: *BEE1* (At1G14800; d*N*/d*S* = 1.09), a positive regulator of the shade avoidance response (Cifuentes-Esquivel, et al. 2013), and *PAR1* (At2G42870; d*N*/d*S* = 1.39), a negative regulator of shade avoidance responses (Bou-Torrent, et al. 2008). These findings suggest that the shade avoidance regulatory pathway may be experiencing relaxed negative selection, perhaps as a result of specialization to open, sparsely vegetated habitats.

Evidence for selection from enrichment analysis

When positive selection is the source of elevated d*N*/d*S*, certain categories of genes — those that include the targets of positive selection — are expected to be statistically over-represented among loci with the highest d*N*/d*S*. Based on this premise, we considered highly significant enrichment of certain gene ontology (GO) categories in the highest d*N*/d*S* category (>1.2) to be a heuristic to aid in the detection of functional categories under positive selection.

From the 1,041 TOPS with d*N*/d*S* >1.2, several classes of GO terms were significantly enriched using Fisher's Exact Test at an FDR <0.05; these categories were dominated by transcription factors and signal transduction pathway molecules, such as kinases and phosphatases (fig. 7). Conversely, other GO categories showed highly significant enrichment in the set of genes with the lowest d*N*/d*S* ratios (0.00–0.099), indicating that purifying selection has been active. This low d*N*/d*S* gene set was dominated by catabolic and anabolic enzymatic functions. These results indicate that, despite small population sizes, natural selection has remained active in these two taxa, and that both positive and negative selection likely played discernable roles in the overall distribution of d*N*/d*S*.

Further, several classes of GO terms were significantly enriched (FDR<0.05) in the 549 TOPs with ω_s >1.2 (when compared to the 7,252 TOPs in which d*S* = 0 and ω_s <1.2; table 3). Some of these enriched GO terms, including ‘transcription factor’ (GO:0003700) and ‘DNA binding’ (GO:0003677), were identical to those enriched in the d*N*/d*S*>1.2 category, despite the fact that these enrichment analyses were based on completely non-overlapping sets of orthologous genes. This recurrence of the same significantly enriched terms in loci with both d*N*/d*S*>1.2 and ω_s >1.2 strongly supports a hypothesis that these GO categories are evolving under positive selection.

Term	GO-ID	0-0.09 (11810)	0.1-0.29 (10726)	0.3-0.79 (11731)	0.8-1.19 (2638)	>1.2 (2156)
Phosphorelay response	GO:0000156					
Nucleotide binding	GO:0000166					
Nucleic acid binding	GO:0003676					
DNA binding	GO:0003677					
Transcription factor	GO:0003700					
Aminopeptidase activity	GO:0004177					
Triglyceride lipase activity	GO:0004806					
Cyclin-dependent inhibitor	GO:0004861					
Sulfate transporter	GO:0008271					
Ran GTPase binding	GO:0008536					
Oxidoreductase activity	GO:0016701					
Hydrolase activity, ester bonds	GO:0016788					
Hydrolase activity, C, N	GO:0016811					
Oxidoreductase activity	GO:0016863					
Ras GTPase binding	GO:0017016					
Purine nucleotide binding	GO:0017076					
Carbohydrate phosphatase	GO:0019203					
SUMO enzyme activity	GO:0019948					
Ribonucleoside binding	GO:0032549					
Ribonucleotide binding	GO:0032553					
Anion binding	GO:0043168					
ADP binding	GO:0043531					
GTPase binding	GO:0051020					
Carbohydrate binding	GO:0097367					
Nucleoside phosphate binding	GO:1901265					
Sulfur compound transporter	GO:1901682					

Figure 7. Heat map of enriched GO categories for TOPs based on d*N*/d*S* values. Number of gene pairs in each category is indicated in parentheses. Colors indicate FDR values, light grey = 0.05, grey = 0.01, and black = 0.005.

PHL1 as a candidate gene for tolerance to limiting phosphate

Serpentine soils have long been known to be moderately to severely deficient in P (Whittaker 1954b). Therefore, we expected *Cab* plants to have evolved molecular mechanisms to deal with continual phosphate limitation. An ortholog of the MYB transcription factor *PHL1* (*PHR-like 1*)

Table 3. List of enriched GO terms in transcripts with a synthetic d*N*/d*S* > 1.2. Default FDR of 0.05 was implemented.

GO-ID	Term	Category
GO:0005634	nucleus	C
GO:0006355	regulation of transcription, DNA-templated	P
GO:0003700	transcription factor activity, sequence-specific DNA binding	F
GO:0010165	response to X-ray	P
GO:0003677	DNA binding	F
GO:0000103	sulfate assimilation	P

was identified as the transcription factor with a very high d*N*/d*S* value (1.79) between Caa and Cab orthologs within a 405 bp alignment that covered a portion of the ~1.1kb CDS (370 aa ORF) predicted from *Ath* ortholog *PHL1* (At5G29000). *PHL1* is a closely related paralog of the *PHR1*, a key regulator of phosphate starvation responses (Rubio, et al. 2001). *PHL1* has a discernable, yet poorly understood role in phosphate starvation responses (PSR) in plants (Bustos, et al. 2010). Targeted mapping of CAA1 and CAB1 reads to the *Ath* CDS database yielded assembled transcripts orthologous to *Ath PHL1* with full-length ORFs of 370 aa and 385 aa, respectively. In CAB1, a 45 bp (15 aa) in-frame insertion adjacent to the second helix of the MYB DNA-binding site (fig. 8a) is expected to completely disrupt the specific DNA binding activity of this transcription factor, implying that *Cab* has a functionally null allele at this locus.

To follow up on this finding, we examined the phenotype of an *Ath* mutant homozygous for a null allele of *PHL1* (SALK_079505C) obtained from the SALK collection of indexed T-DNA insertion lines (Alonso, et al. 2003). As shown in figs. 8b and 8c, we observed superior growth in limiting phosphate in both *Cab* (relative to *Caa; P* = 0.0001 at 10 µM PO_4^{2-}) and in the *Ath phl1* knockout line (relative to *Ath* wild-type Col-0; *P* = 0.001 at 40 µM PO_4^{2-}). Thus, in very low phosphate conditions, we observed an approximately two-fold growth advantage, measured

in terms of dry biomass, in the *phl1* loss-of-function *Ath* mutant (relative to wild type) and in *Cab* (relative to *Caa*) suggesting that loss-of-function alleles in *PHL1* may have an adaptive role in the phosphate-limited conditions.

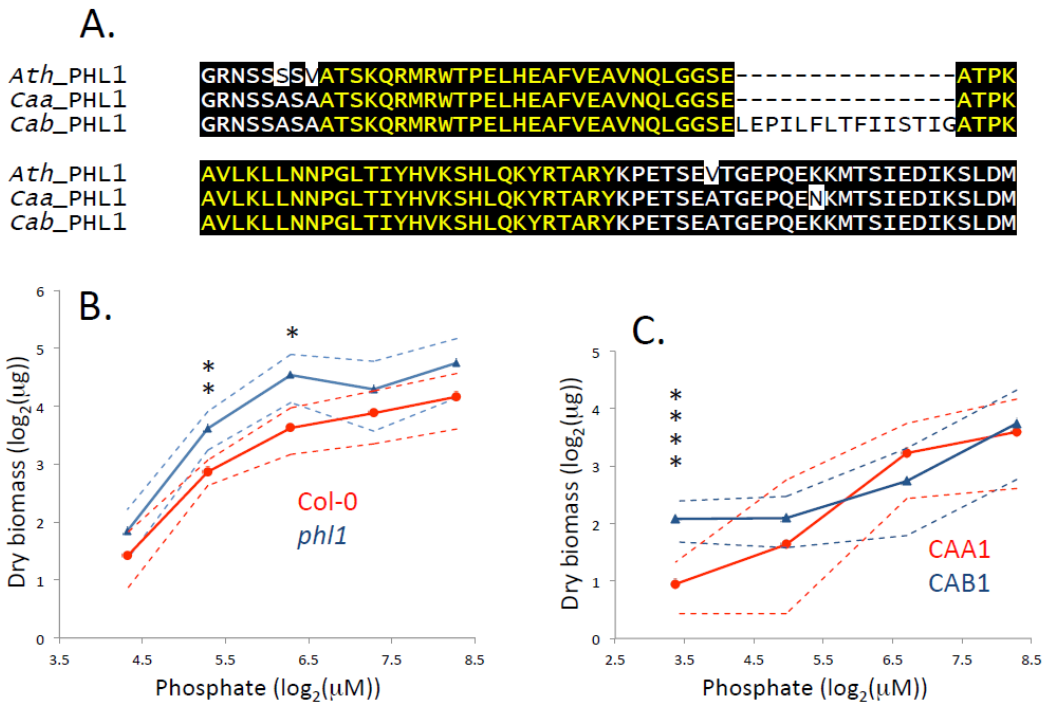

Figure 8. Phenotypes of variants in the MYB-CC transcription factor *PHL1*. a Protein alignment of *Ath*, *Caa*, and *Cab* PHL1. Yellow font indicates the conserved MYB DNA-binding domain. **b** Seedling growth phenotypes of *Ath* wild-type (Col-0) and *phl1* mutant in deplete phosphate conditions. Growth metric is mean end-point dry biomass measured 22 d after emergence of first true leaf. Dotted lines (----) indicate +/- one standard deviation from the mean. Significance in 2-tail t-tests is indicated as * =0.05, ** =0.01, **** =0.001. **c** Seedling growth phenotypes of CAA1 and CAB1 in low phosphate conditions. Growth metric is end-point dry biomass measured 28 d after emergence of first true leaf.

GO Enrichment of loci with nonsense and indel polymorphisms

TOPs with stop codon polymorphisms and those with indels (with or without putative frameshifts) were not used in PAML analyses. However, analyses of such transcripts could

reveal biologically important differences in gene function between the *Caa* and *Cab*. GO enrichment analyses were performed comparing the 241 TOPs with indels against all TOPs without indels; There were no significantly enriched GO terms found (at FDR <0.05). However, there were several GO terms that were significantly enriched in the set of 284 TOPs with stop-codon polymorphisms. These included 'inositol trisphosphate kinase activity' (GO:0051765) and 'sulfate assimilation' (GO:0000103) (supplemental tables 4C&D, attached files).

Evolutionary divergence of Caa and Cab

To better understand the evolutionary history of *Caa* and *Cab*, synonymous substitution rate (d*S*) was used to estimate the relative timing of divergence events in both the nuclear and chloroplast genomes. These d*S* values indicated a very recent divergence of chloroplast orthologs (mean d*S* = 8.4 x 10^{-5}, range = 0.0000–0.0025). Using a comparison of CAA1 and *A. thaliana* plastid and nuclear orthologs, we observed that the nuclear synonymous substitution rate within the clade was ~3.9 fold higher than that of the plastid genome. We used this higher nuclear substitution rate to adjust the plastid d*S* values in order to make meaningful comparisons with the nuclear genome divergence, resulting in an adjusted mean d*S* = 0.0003 with a range = 0.000–0.009 in the plastid-derived transcripts.

The d*S* values of nuclear orthologs had a much broader distribution (mean d*S* = 0.021, range = 0.00-3.87) with two highly distinct peaks (figs. 9, 10). The nuclear d*S* distribution had a bimodality coefficient of 0.677 (a value of >0.55 indicates the data is bimodal, and a value of 1.0 is only obtained when the dataset consists of two separate point masses). The bimodality amplitude, in which larger values indicate more distinct peaks, was 0.996 (where 1.0 corresponds to two separate point masses).

The extremely high resolution afforded by deep next-generation sequencing can be used to support a hypothesis of introgression as opposed to alternative explanations such as incomplete lineage sorting or highly variable mutation rates for individual genes. Here, introgression is supported by a distinct peak of alleles that diverged long after the initial point of divergence (Twyford and Ennos 2012; Brandvain, et al. 2014). In our case the characteristics of the peak with the lower mode of d*S* are inconsistent with both the alternative hypothesis of variation in mutation rate, and that of lineage sorting. Specifically, with substantial variation in nuclear mutation rate, we would expect a very broad secondary peak, or perhaps a left-skewed tail of the main peak of d*S* values. The observed compactness (low variance) in the peak of lower d*S* is thus inconsistent with global variation in mutation rate (unless the nuclear genome is somehow partitioned into two gene sets with extremely different mutation rates). With lineage sorting, we would expect to see a left-skewed tail of d*S* values arising from the main peak of divergence. Instead we see a vastly separated peak of seemingly recent sequence divergence that is consistent with recent introgression. The discordance between nuclear and plastid divergence times also supports the hypothesis of introgression (Twyford and Ennos 2012). Thus, the strongly bimodal distribution of d*S* values with a large separation between peaks (fig. 9), suggested two separate divergence events (Twyford and Ennos 2012; Brandvain, et al. 2014), with the left-most peak representing 9,871 orthologs (~33.2% of total) that diverged after very recent hybridization and introgression, and the right-most peak (modal d*S* = 0.026) indicative of a much more ancient divergence event involving 19,472 orthologs (~66.1%).

BEAST analysis of a set of 56 plastid-encoded genes yielded an estimated divergence date of 0.125 Ma (95% c.i. of 0.027 Ma to 0.232 Ma), while analysis of alignments of 24 randomly selected nuclear genes yielded a divergence time of 2.8 Ma (95% c.i. 2.3 Ma to 3.2

Ma). Importantly, this global estimate of nuclear divergence time does not take into account the strongly bimodal distribution of divergence of individual genes, with the majority falling into the peak corresponding to earlier divergence. Indeed, BEAST analysis of a set of 50 genes from near the modal d*S* value of the putative earlier divergence (d*S* = 0.025-0.027) yielded an estimated divergence time of 3.2 Ma (95% c.i. of 2.8 Ma to 3.6 Ma). From these results, we surmised that an initial divergence event occurred ~3 Ma and that secondary contact and introgression, revealed by both the chloroplast genes and a minority of nuclear genes, occurred relatively recently (~0.12 Ma).

Recent gene duplication

All-against-all BLASTN analyses of CAA1 vs. CAA1 and CAB1 vs. CAB1 was used to identify 1,299 and 729 high-confidence tentative paralogous pairs (TPPs), respectively (supplemental table 3, in Appendix A). These TPPs had a mean pairwise d*N*/d*S* ratio of 0.1917 (median = 0.123) in CAA1 and 0.1989 (median = 0.130) in CAB1. Since both duplicates continue to be transcribed and the mean d*N*/d*S* ratios between paralogs were both significantly lower than the global mean between orthologs ($P<0.0001$), we concluded that, in this set of genes, negative selection is acting to retain both gene copies in a functionally active state.

Distributions of d*S* values for TPPs were graphed (fig. 10) to identify peaks corresponding to episodes of elevated gene duplication (Blanc and Wolfe 2004). In both taxa, peaks of duplication at the far-left (d*S* = 0.000–0.005) provide evidence of very recent gene duplications; These left-edge peaks are commonly observed and presumably arise from background segmental (e.g., tandem) duplication (Blanc and Wolfe 2004). We also observed

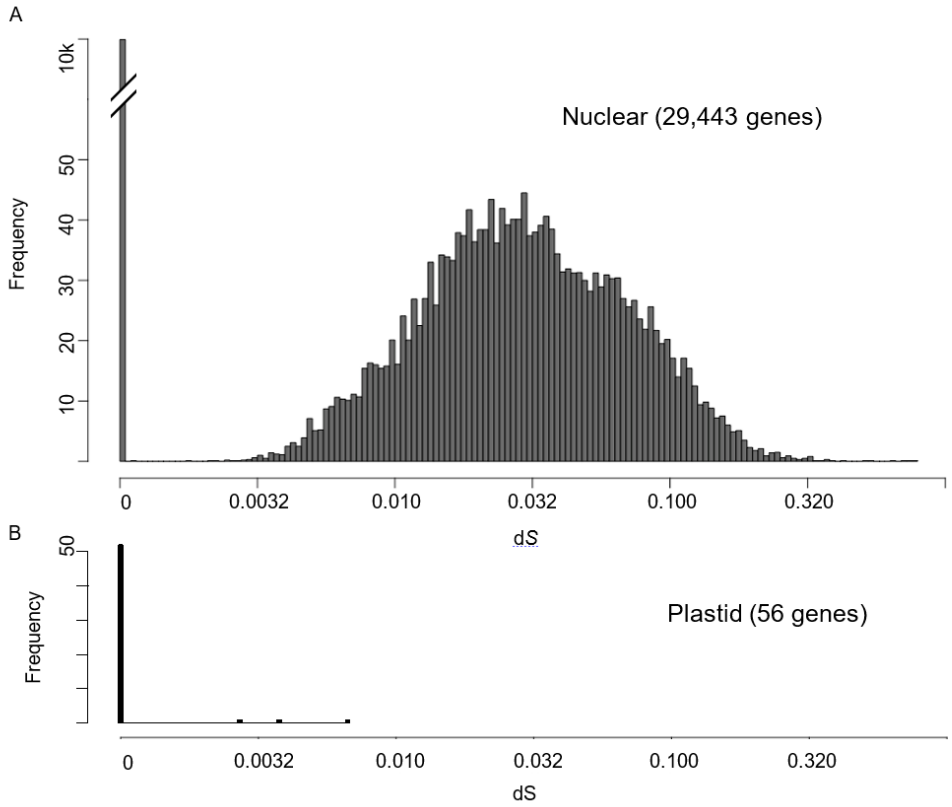

Figure 9. Rates of d*S* between CAA1 and CAB1 orthologs. **A)** Histogram of d*S* values for 29,443 nuclear genes. Column at the far left is comprised of orthologous pairs with d*S* = 0. **B)** Histogram of d*S* values for 56 plastid encoded genes. Column at the far left corresponds to orthologous pairs with d*S* = 0. Non-zero plastid d*S* values were adjusted upward (as described in Materials and Methods) to compensate for a faster nuclear divergence rate, and thus allow for a relevant comparison with nuclear d*S* values.

evidence for earlier gene duplication, some of which can be explained by the Brassicaceae α whole-genome duplication event (Initiative 2000; Bowers, et al. 2003) that occurred prior to the major divergence of Lineage I and Lineage II of the Brassicaceae family (fig. 10a) and is characterized by paralogs with a mean d*S* of ~0.8 (Guo, et al. 2013; Kagale, et al. 2014). Our dataset showed additional duplication events that, based on d*S* values, took place in the interval after the divergence of Lineage I and Lineage II (~23.4 Ma; fig. 10b), but prior to the initial episode of nuclear divergence of *Caa* and *Cab* (~3 Ma). The modes of this peak of duplication

were not well defined in either *Caa* or *Cab*, but were roughly in the interval of d*S* values between 0.08 and 0.16 (fig. 10b).

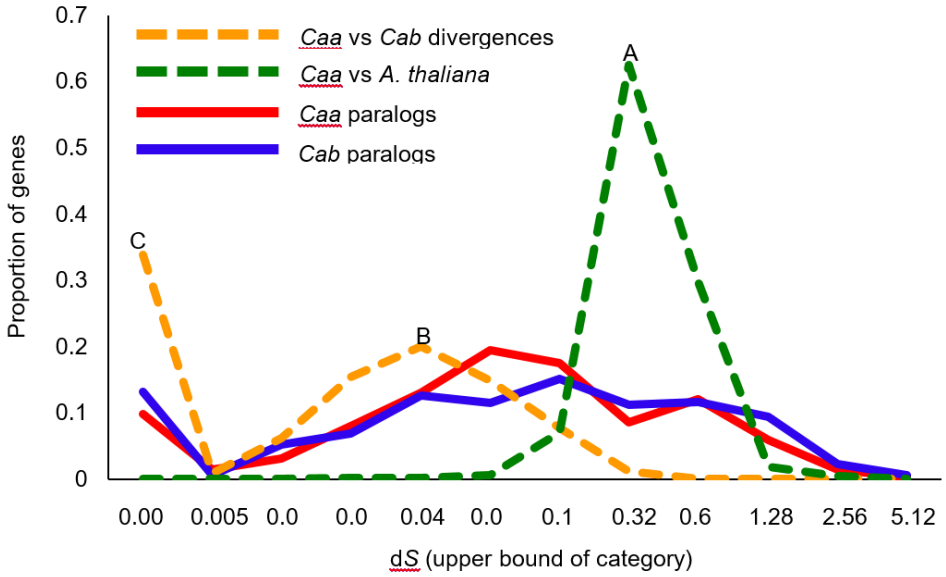

Figure 10. d*S* comparisons between orthologs (dashes) and paralogs (solids). A denotes the d*S* values of *Caa* vs. *Ath* orthologs, with the major peak presumably corresponding to the Brassicaceae Lineage I/Lineage II (*Caa*/*Ath*) split ~23.4 Ma. **B** denotes d*S* values of Caa vs. Cab orthologs, with the right-most peak corresponding to the putative earlier divergence between *Caa* and *Cab* lineages (~3.2 Ma). **C** denotes the much more recent divergence of *Caa* from *Cab* orthologs (~0.125 Ma).

Of the TPPs in this d*S* interval, 194 pairs were retained in both CAA1 and CAB1. Based on parsimonious consideration of the d*S* values, along with the shared nature of these TPPs, it is likely that the bulk of these genes duplicated prior to the first divergence of the *Caa* and *Cab* lineages. To compare the evolutionary outcomes of these newly duplicated genes in the two separate lineages, d*N*/d*S* distributions for the paralogs were compared between CAA1 and CAB1 (fig. 11). These paralog sets had had similar mean d*S* values (0.177 in CAA1 and 0.183 in CAB1) and the overall distributions of d*N*/d*S* were very similar (fig. 11b), with very few genes

exhibiting neutral evolution (i.e. d*N*/d*S*~1) in either taxon (fig 11a). Thus, in this gene set, negative selection appears to have acted to retain two functional gene copies in both taxa.

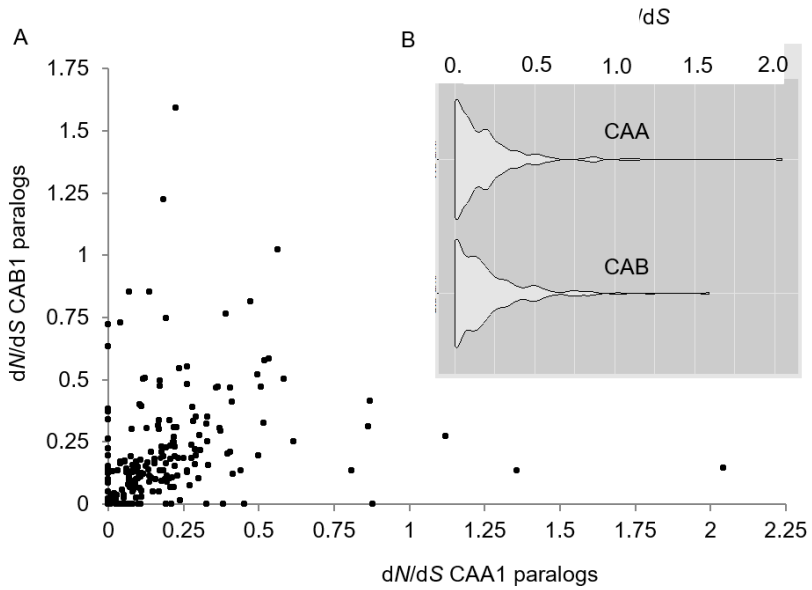

Figure 11. **Coding sequence evolution in paralogs shared between CAA1 and CAB1. A.** a scatterplot of d*N*/d*S* values of 194 corresponding parologous pairs. **B.** density distribution (violin) plots of d*N*/d*S* values for 194 parologous pairs within the CAA1 and CAB1 transcriptomes.

Discussion

Serpentine barrens are an extremely harsh environment that presents multiple chemical and physical challenges to plant life. These include toxic heavy metals, limiting mineral nutrients, water stress, high light levels, and temperature stress. This study describes the development and annotation of comprehensive reference transcriptomes for serpentine tolerant *Cab* compared to its serpentine intolerant sister taxon *Caa.* While this work focused on the analysis of coding sequence evolution, we also obtained several unexpected but critical insights into the evolutionary-genomic histories of these two highly ecologically divergent taxa.

From our merged assemblies, we observed ~94,000 representative transcripts from *Caa* and ~83,000 transcripts from *Cab*. This difference in transcript number likely reflects the greater depth of sequencing in CAA1 (~80 million vs. ~50 million reads) as well as the lack of representation of floral and fruit tissues in the CAB1 RNA pool. Although the overwhelming majority of these representative transcripts had significant BLAST hits to the NCBI nr and nt databases (at threshold *e*-values of $<10^{-6}$ and $<10^{-8}$ respectively), this finding does not imply that all such transcripts correspond to genes encoding functional proteins or RNAs. Pervasive 'background' transcription of the nuclear genome is a well-known phenomenon in many species (Jensen, et al. 2013; Neme and Tautz 2016). Whether these pervasive transcripts are functional or merely 'spurious' is a subject of vigorous debate (Graur, et al. 2013). In this work, we focused on a set of ~29,000 genes that have orthologs in CAA1 and CAB1 based a reciprocal best-BLAST hit relationship, that are thus subject to meaningful direct comparisons between the two taxa.

Multiple factors affect* d*N*/d*S* *ratio

Beyond positive selection, there are a number of explanatory mechanisms for the relatively high d*N*/d*S* ratio we observed among orthologs in CAA1 and CAB1. One of these is the intrinsic nonlinearity in the accumulation of mutations. Among close evolutionary relatives, this can lead to ortholog pairs with high d*N*/d*S* by chance due to stochasticity of the mutation process, along with limited divergence time (e.g., d*N*>0 and d*S*~0)(Wolf, et al. 2009; Montoya-Burgos 2011; Mugal, et al. 2014). A further possible explanation of elevated d*N*/d*S* is relaxation of negative selection on certain ecologically relevant genes when an organism colonizes a novel environment. In this scenario, some genes may become dispensable (Lahti, et al. 2009) and show

a trend toward neutrality (e.g., d*N*/d*S* ~1). Additionally, genetic drift can reduce or nullify the effects of both positive and negative selection. For example, a genome-wide decrease in the efficacy of selection against weakly or moderately deleterious alleles results in an elevated mean d*N*/d*S* (Wright 1931a; Ohta 1972; Kimura 1984; Charlesworth 2009; Strasburg, et al. 2011).

In this study, we set out to determine if any of these mechanisms played a demonstrable explanatory role in the observed elevated transcriptome-wide mean d*N*/d*S* ratio. A small number of loci with very small d*S* values did indeed appear on the roster of genes with high d*N*/d*S* ratios. To determine the effect of these genes on our global estimate of d*N*/d*S*, we recalculated this parameter in the subset TOPs with the arbitrarily defined threshold value of d*S* >0.04. The 5,878 TOPs in this subset had a mean d*N*/d*S* of 0.343 and a median of 0.248 — nearly identical to our global mean and median values of 0.346 and 0.220, respectively. From this comparison, we concluded that the few loci that had a high d*N*/d*S* due to very low d*S* made little contribution to our genome-wide d*N*/d*S* mean.

To test for the effect of genetic drift resulting from small population size, d*N*/d*S* ratios between *Caa* and *Cab* were compared with ratios between *Capsella grandiflora* and *Arabidopsis lyrata*, which both have relatively large population sizes (Slotte, et al. 2010). In this set of 177 orthogroups (Slotte, et al. 2010) we observed a significantly higher d*N*/d*S* in the *Caa* vs. *Cab* comparison across the entire spectrum of d*N*/d*S* ratios ($P = 0.001$). Since the gene set was composed of loci that were chosen without regard to function, they would not be expected to be broadly subjected to differential selection (positive or negative) in one species pair versus the other. Rather, this comparison implied the existence of a discernable effect of small population size on the observed high d*N*/d*S* ratios in *Caa* vs. *Cab.*

We also examined d*N*/d*S* ratios for genes in the plant 'shade avoidance' GO category, which we anticipated might have a diminished fitness benefit in the sparsely vegetated serpentine barrens and granite outcrops of *Cab* and *Caa* respectively. After removal of four highly pleiotropic genes, we observed an elevated d*N*/d*S* = 0.470 in this category, with two shade avoidance regulators apparently evolving under neutrality. These findings suggest that the shade avoidance regulatory pathway may be experiencing relaxed selection as a result of specialization to open, sparsely vegetated habitats. This hypothesis can be readily tested by quantifying shade avoidance responses in *Caa* and *Cab* in comparison to less edaphically specialized species in the Brassicaceae family (such as *Athal*). Importantly, relaxed negative selection offers a compelling genetic explanation for the endemic restriction of these taxa, as they would be poorly adapted to more favorable environments that would support greater vegetation densities and thus greater competition.

Considered together, our observations indicate that the elevated global d*N*/d*S* ratio we observed is likely the result of a composite of explanatory processes including positive selection, genetic drift, and relaxation of negative selection. This work provides a useful example of the effects of multiple factors d*N*/d*S* ratio, with the implication that this metric should only be used (with caution) as an early exploratory heuristic for identifying candidate genes under positive selection—particularly in taxa where N_e is small or not known.

Biological insights from coding sequence evolution

Genome ontology (GO) annotation of the gene-set with the highest d*N*/d*S* (>1.2) yielded a broad distribution of molecular functions, including several categories that might be involved in serpentine tolerance such as 'Ion Binding '(299 genes), 'Transcription Factors' (96 genes),

'Oxidoreductase Activity' (64 genes), and 'Transmembrane Transporter Activity' (33 genes; fig.5). These categories included orthologs of a number of *A. thaliana* genes with known functions in mineral nutrition and heavy metal transport such as phosphate transporter *PHT4;4* (AT4G00370.1, *e*-value = $8.57E^{-33}$, d*N*/d*S* = 1.75), potassium transporter *KT1* (AT2G26650.1, *e*-value = 0.0, d*N*/d*S* = 1.56), and the heavy metal transporter *HMA2* (AT4G30110, *e*-value = $3.2E^{-26}$, d*N*/d*S* = 1.32), in which metal binding specificity is determined largely by the amino acid sequence of the N-terminal domain (Zimmermann, et al. 2009). As a category, transcription factors were highly enriched in the set of orthologous pairs with the highest d*N*/d*S* values (>1.2) and in the set with the highest ω_s (>1.2)

Similarly, GO terms related to sulfate assimilation and transport (GO:0000103, GO:1901682 and GO:0008271) were significantly enriched in genes with d*N*/d*S* from 0.8 to 1.19, ω_s >1.2, and those loci with stop-codon polymorphisms (fig. 7, table 3, supplementary table 4B, attached file), implying that evolution of sulfate transport pathways may have a discernable role in serpentine adaptation. The gene set with d*N*/d*S* from 0.8 to 1.19, which could include both genes evolving under neutrality and those in which only a subset of residues are evolving under positive selection, includes orthologs of four *Athal* genes annotated as sulfate transporters: *SULTR1;3* (AT1G22150, d*N*/d*S* = 0.89), *SULTR2;2* (AT1G77990, d*N*/d*S* = 0.90), *SULTR3;2* (AT4G027003;2, d*N*/d*S* = 1.12) and *SULTR3;3* (AT1G23090, d*N*/d*S* = 0.89). In this regard, it is important to note that another sulfate transporter, *SULTR1;1*, was implicated as a candidate locus that was subjected to a selective sweep during adaptation to serpentine in *A. arenosa* (Arnold et al. 2016).

There is substantial disagreement in the literature as to whether sulfur is more limiting in serpentine than other soil types. In some annual grasses, addition of nitrogen and phosphorous

had the biggest effect on increasing biomass production in serpentine soils, while sulfate addition had little effect (Turitzin 1982). In contrast, strong responses to the addition of both phosphorous and sulfate were observed in subterranean clover (Jones, et al. 1977). Few, if any, experiments have been performed that directly determine the degree of sulfur limitation in serpentine soils. The recent genomic studies from our laboratory and others provide compelling examples of the 'reverse ecology' approach (Ellison, et al. 2011) by suggesting that genes involved in sulfur transport may play an important role in adaptation to serpentine, and justify further investigation of the role of sulfur limitation in the serpentine environment.

In contrast to sulfur, phosphorous deficiency is a well-established attribute of serpentine-derived soils (Whittaker 1954b). One of the genes implicated by a high pairwise d*N*/d*S* ratio between CAA1 and CAB1 was the transcription factor *PHL1*, which plays a known, but poorly understood role in responses to phosphate deficiency. Detailed investigation of this transcript showed that the CAB1 allele has a 15 aa insertion within the DNA binding domain that would eliminate its DNA binding capacity. We further showed that an *Athal* line homozygous for a *phl1* loss of function allele showed superior growth responses in extremely low phosphate conditions. These observations are consistent with a scenario in which positive selection has favored alleles with reduced gene function at the *PHL1* locus, as has been observed in a number of other cases of adaptive evolution such as the *pitX1* and *eda* loci of stickleback, and the *CCR5* and *DUFFY* loci in humans (Mummidi, et al. 1998; Hamblin, et al. 2002; Shapiro, et al. 2004; Colosimo, et al. 2005).

Evolutionary implications of reticulate evolution

Population-genomic investigation of a serpentine-tolerant population of *Arabidopsis arenosa* indicated that a subset of putative serpentine-adaptive alleles, identified by signatures of selective sweeps, appear to have been recently introgressed from *Arabidopsis lyrata* (which diverged from *A. arenosa* ~0.4 Ma)(Arnold, et al. 2016). In our study, we also found evidence for reticulate evolution, in that most of the nuclear genes (~66%) in *Caa* and *Cab* lineages are estimated to have diverged ~3 Ma, and that a minority of genes (~34%), along with the entire plastid genome, diverged much more recently (~0.12 ma). The areas of serpentine in the San Rafael Mountains that are presently habitat to *Cab* became exposed during the middle Pliocene to early Quaternary (~1.0–3.5 Ma) (Raven and Axelrod 1995). During this period, southern California was subject to large climatic changes (Raven and Axelrod 1995). Although *Caa* and *Cab* are geographically isolated at present, periods of wetter and cooler climatic conditions in the past may have led to reduced competition, and expansion of populations, facilitating secondary contact and resulting in the observed pattern of introgression.

The 'Streptanthoid Complex' includes the non-monophyletic genera of *Caulanthus* and *Streptanthus* (Burrell, Taylor, et al. 2011; Ivalu Cacho, et al. 2014), and comprises much of the North American species in the tribe Thelypodieae. Present day species from throughout the Streptanthoid Complex are largely interfertile in experimental crosses (A.M. Burrell unpublished; K. Christie et al., unpublished). Our finding of likely secondary contact and hybridization within the *Caulanthus amplexicaulis* lineage several million years after initial divergence reveals the possibility that reticulate evolution may have played a role in the attainment of serpentine tolerance across the broader Streptanthoid Complex and Thelypodieae tribe.

The Streptanthoid Complex and the encompassing Thelypodieae tribe have been highly recalcitrant to attempts at phylogenetic reconstruction. The resulting trees have been characterized by large unresolved polytomies (Pepper and Norwood 2001; Warwick, et al. 2009; Mayer and Beseda 2010). In the best-resolved phylogeny (Ivalu Cacho, et al. 2014), six nuclear markers and two plastid markers were used to estimate that serpentine tolerance arose independently ~4 times in the "ASHTB" sub-clade of the complex (credibility = 0.97) that includes *C. amplexicaulis*, the Sierra Nevada serpentine endemic *Streptanthus polygaloides, S. tortuosus* (which has serpentine tolerant and non-tolerant ecotypes), a number of serpentine-tolerant *Streptanthus* species of the California Coastal Range, and several non-serpentine species. The Sierra Nevada and Coast Range outcrops on which the serpentine species are presently found probably became exposed nearly contemporaneously during the Pliocene (2.58–5.33 Ma) or later (Raven and Axelrod 1995), a plausible time of divergence for the ASHTB clade. These findings reveal the possibility that introgression among lineages may have played a role in the pattern of serpentine tolerance in this clade. Specifically, the apparently separate gains of serpentine tolerance (Ivalu Cacho, et al. 2014) may not have been independent at the level of individual alleles (i.e., independent clades may share adaptive alleles that are identical by descent) — a finding that would dramatically alter our model for the evolutionary acquisition of serpentine tolerance in this group of taxa

Implications of recent gene duplication

Gene duplication, along with subsequent neofunctionalization, subfunctionalization, and changes in gene dosage and expression, have long been considered to play important roles in evolutionary adaptation to novel environments (e.g., Ohno 1970; Lynch and Conery 2000; Zhang 2003;

Hughes 2005; Conant and Wolfe 2008; Flagel and Wendel 2009; Kondrashov 2012; Makino and Kawata 2012; Tamate, et al. 2014; Schlötterer 2015; Loehlin and Carroll 2016). In plants, gene duplication has been recognized as an important component of adaptive tolerance to cadmium, zinc, aluminum, and sodium ions (Dassanayake, et al. 2011; Craciun, et al. 2012; Oh, et al. 2014). Our analysis of d*S* values between paralogs within CAA1 and CAB1 yielded evidence of elevated gene duplication in the shared ancestral lineage of *Caa* and *Cab*, with diffuse peaks of genes with d*S* values in the interval between 0.08 and 0.16 (fig. 10). Interestingly, *Pringlea antiscorbutica* and *Stanleya pinnata*, which are also members of the Thelypodiae tribe, show signatures of elevated gene duplication with a peak of d*S* values from 0.12 to 0.19, with an estimated time of divergence of ~10 Ma (Kagale, et al. 2014). Recently, signatures of a similar duplication event have been observed in *Streptanthus farnsworthianus* using whole transcriptome sequencing and comparative chromosome painting (Mandáková et al. 2017). Most parsimonious phylogenetic placement of events indicates that these gene duplications occurred in a common ancestor to *Pringlea*, *Stanleya*, *Streptanthus*, and *Caulanthus* (Bartish, et al. 2012; Ivalu Cacho, et al. 2014; Kagale, et al. 2014). The tribe Thelypodiae is noted for its unusually high levels of both ecological and morphological diversity (Al-Shehbaz, et al. 2006; Burrell, Taylor, et al. 2011; Ivalu Cacho, et al. 2014). For example, *Pringlea* is a monotypic genus that is adapted to frigid islands of the sub-Antarctic ocean and has a woody tree-like morphology (Bartish, et al. 2012). *Stanleya pinnata* has a highly unusual floral structure and has edaphic ecotypes that are adapted to high sodium, boron, and selenium in soils (Feist and Parker 2001; Freeman and Banuelos 2011). Several other species in the Thelypodiae tribe show remarkable morphological diversity as well as adaptation to wide range of difficult edaphic environments including serpentine (Pepper and Norwood 2001; Burrell, Taylor, et al. 2011; Ivalu Cacho, et al.

2014). It is tempting to speculate that a duplication event (e.g., whole genome duplication) in the lineage leading to the Thelypodieae may have contributed to the morphological and ecological diversity of this tribe.

The 194 paralogous gene pairs that are shared in both CAA1 and CAB1 show a similar pattern of d*N*/d*S* that is consistent with negative selection acting to maintain duplicate functional copies of these genes. The retention of these gene pairs could reflect exposure to a common set of selective pressures in both lineages. Species in the Thelypodiae tribe are adapted to open, rocky habitats, and it has been hypothesized that adaptation to such environments may be a prerequisite trait for colonization on serpentine (Pepper and Norwood 2001; Ivalu Cacho, et al. 2014). Gene duplication may have played a role in this adaptation, as several duplicated genes that are shared between *Caa* and *Cab* have established roles in heat and water stress. These include orthologs of *A. thaliana* genes *OCP3* (At5G11270), a homeodomain transcription factor involved in drought tolerance (Ramírez, et al. 2009) and *CRT3* (At1G08450), a calreticulin with roles in calcium ion homeostasis and tolerance to water stress (Christensen, et al. 2010).

Reconciling selection and drift

Sewell Wright suggested that selection becomes ineffective when $N_e s<1$ (Wright 1931b). In the case of both Caa and Cab we observed very small present-day N_e values, yet found compelling evidence for both positive and negative selection. Given the fundamental constraint described by Wright, it is difficult to support a model in which key events in the evolution of serpentine tolerance occurred either by 1) selection acting on a large number of loci of small effects (i.e., small *s*) or 2) in the context of very small populations (i.e., small N_e). Rather, population sizes may have been much larger during key episodes of adaptive evolution. Other serpentine

endemics in the Thelyopdiae tribe are sometimes found in larger populations than *Cab*. For example, various subspecies of the serpentine endemic *Streptanthus morrisonii* have population sizes of 100–2,000 (Dolan 1995). *Streptanthus niger*, another serpentine endemic, occurs in populations as large as 4,000-8,000 individuals (Sarah Swope, personal communication). Thus, it is plausible that at some point in the past, climatic or edaphic conditions supported larger ancestral populations on or near serpentine outcrops. In an alternative scenario, allelic differences with large enough values of *s* can still undergo selection even in small populations (i.e., $N_e s > 1$). In such populations, dramatic evolutionary changes would be expected to occur through very few loci that have large effects. For example, our previous QTL mapping of nickel tolerance in the F_2 progeny of a cross between *Caa* and *Cab* uncovered two loci with large effects that explained 28% and 26% of the total phenotypic variance, respectively (Burrell, et al. 2012). These two scenarios that explain adaptive evolution despite small contemporary population sizes are not mutually exclusive.

Summary and perspectives

We obtained a comprehensive catalog of coding sequence variants between two ecologically divergent plant varieties, one of which is tolerant of — and endemic to — the difficult serpentine geological environment. We examined for signatures of positive selection in ortholog pairs with a high d*N*/d*S* ratio, but demonstrated that, in our study taxa elevated d*N*/d*S* ratios were likely the result of several distinct and sometimes antagonistic evolutionary processes. For any given locus, disentangling these effects remains a considerable challenge. However, our characterization of d*N*/d*S* and a novel synthetic d*N*/d*S* metric, along with patterns of coding sequence evolution that is being used, along with other sources of evidence, including QTL mapping in CAA1 x CAB1

crosses (Burrell, et al. 2012), population genomics, and gene expression studies, to identify high-confidence candidates for genes underlying serpentine tolerance. Importantly, this work led us to reject a simple model of differential adaptation on a path to bifurcating speciation (cladogenesis driven by divergent environmental pressures) for a more complex but better resolved evolutionary history that incorporates genetic drift, relaxed selection, gene duplication and introgression.

CHAPTER III

RNA-SEQ ANALYSIS REVEALS NOVEL INSIGHTS IN SERPENTINE TOLERANCE IN *CAULANTHUS AMPLEXICAULIS* VAR. *BARBARAE* (BRASSICACEAE).

Overview

Serpentine endemic plants are excellent models for the study of molecular evolution as they provide extreme examples of adaptation to environment. Serpentine outcrops are derived from ultramafic rock and have extremely low levels of essential plant nutrients (e.g. N, P, K, Ca), as well as toxic levels of heavy metals (e.g. Ni, Cr, Co), and very poor moisture availability. These outcrops provide habitat to the endemic plant species, *Caulanthus amplexicaulis* var. *barbarae* (J. Howell) Munz ("*Cab"*). Its sister species, *C. amplexicaulis* var. *amplexicaulis* S. Watson (*Caa*), is found predominately on granite soils and is intolerant to serpentine soils. The objective of this project was to answer the following three questions: (1) which genes are constitutively expressed in *Cab*, (2) which genes are induced in *Cab* on serpentine outcrops, and (3) which genes are induced in *Caa* on serpentine outcrops?

Comprehensive reference transcriptomes and genomes of both *Caa* and *Cab* are available for use in protein coding gene comparisons. Both common-garden on ¼ strength MS media and reciprocal transplant experiments on natural granite and serpentine soils were performed using *Caa* and *Cab*. RNA-seq analyses were implemented to calculate global expression patterns and identify differentially regulated genes that may play a role in serpentine adaptation. Gene expression levels were calculated as transcripts per million (TPM) by the program Salmon using the CAB1 genome assembly as reference and paired end fastq files of RNA-seq data from all replicates in all growth conditions. EdgeR was then used to calculate log fold change for all

genes. Weighted correlation network analysis (WGCNA) was performed to cluster all genes based on their expression profile. In all, the data was separated into 18 modules (clusters) based on similar expression levels. These modules were further processed using fisher's exact test to find any overrepresented GO terms. Gene expression from RNA-seq data implicated a suite of chloroplast and plastid genes that were constitutively expressed in *Cab*; this is an unexpected and novel finding. Genes induced in *Cab* on serpentine include those with roles in nutrient acquisition and movement and heavy metal binding, and genes induced in *Caa* on serpentine indicate response to nutrient starvation and galactose binding/transport. Ultimately results from these analyses, in conjunction with coding sequence data and QTL mapping, will be used to find high quality candidate genes that confer tolerance to serpentine soils.
Key words: adaptation, *Caulanthus*, RNA-seq, serpentine, Thelypodieae, ultramafic

Introduction

Plants are, by and large, immobile organisms unable to relocate when faced with harsh environmental situations. Because of this sedentary lifestyle plants have, arguably, some of the most spectacular examples of adaptations known. Obviously, not all adaptations have resulted from natural selection; however, varying edaphic conditions have long been considered as one of the strongest forces causing evolutionary change (Wallace, 1858).

One of the most arduous environments that land plants have been exposed to and subsequently adapted to, are serpentine outcrops. These outcrops are derived from the alteration of ultramafic rock and are composed of at least 70% mafic minerals, magnesium, or iron (Reeves, et al. 1983; Brady, et al. 2005). They have a world-wide distribution and inhabit almost every type of ecosystem. However, they have a random and patchy distribution with no

intermediate or transition zone. Therefore, adaptation to these outcrops had to be ‘quick’ and most likely include a suite of pleiotropic genes (Pepper, personal communication). Serpentine outcrops are described by a list of unifying features including high rates of endemism, poor productivity from flora and fauna, and unique vegetation from that found in very close proximity (Whittaker 1954a). If given the opportunity, most organisms would opt for another environment to call home; however, refuge from competition is the most well supported hypothesis of why some organisms would occupy these soils.

Serpentine soils are characterized by an amalgamation of extreme environmental conditions. Chemically, these soils have a skewed magnesium to calcium ratio (vastly more magnesium), limiting essential plant nutrients (N, P, K) and excessive heavy metals (Ni, Cr, Cu, Zn, Cd). Physically, the outcrops are usually found on the side of a steep mountain so they are vulnerable to erosion, have shallow soils, and low moisture retention. Biologically, these soils are dark in color and have little plant coverage, which leads to high levels of heat retention and elevated soils temperatures and exposure to high light levels. Few plants have adapted to this lifestyle, but for those that have, they thrive in these conditions. Even more important, many of these plants in serpentine soils have a closely related sister taxa or congener that is unable to live under these extreme conditions. This makes for compelling genetic analyses to uncover the molecular adaptations leading to serpentine survival.

One pair of sister taxa that display this ability vs inability to live in serpentine outcrops is found in the *Caulanthus amplexicaulis* complex (Brassicaceae). *Caulanthus amplexicaulis* var. *barbarae* (J. Howell) Munz, designated *Cab*, is a small annual diploid that is endemic to a small set of isolated outcrops in the San Rafael Mountains of southwest California. Its sister taxon, *C. amplexicaulis* var. *amplexicaulis* S. Watson, designated *Caa*, has a much wider distribution and

is mainly found on granite soils throughout the Transverse Ranges of southern California (Pepper and Norwood 2001). These sister taxa are ecologically and geographically isolated (at their closest they are found separated by ~75km); however CAB and CAA are fully interfertile in artificial crosses (Kruckeberg and Rabinowitz 1985; Pepper and Norwood 2001). They are morphologically very similar with only sepal color as an obvious distinguishing feature. These two taxa are very amenable to work with in a lab setting because they have a very short generation time (seed to seed) of approximately 10-12 weeks, and are both interfertile and self-fertile. Hybrid offspring can be selfed for several generations easily resulting in recombinant inbred lines (RILs) with no major known inbreeding depression issues.

Comprehensive normalized reference transcriptomes have been annotated for both *Caa* and *Cab* (Hawkins et al. 2017). Comparisons of orthologs between *Caa* and *Cab* revealed a high genome-wide d*N*/d*S* ratio between the two taxa (mean = 0.346), which is most likely explained by the composite effects of genetic drift, positive selection, and the relaxation of negative selection (Hawkins, et al. 2017). Further, a period of elevated gene duplication (~10 Ma) was detected by paralog analyses within each taxon, a signature that is shared with other species of the tribe Thelypodieae and may have played a role in the striking morphological and ecological diversity of this tribe. In addition, a strongly bimodal distribution of the synonymous substitution rate, d*S*, was obtained, indicating a history of reticulate evolution and introgression that may have contributed to serpentine adaptation (Hawkins, et al. 2017).

The normalization of these reference transcriptomes was a tradeoff. The reduced representation of coverage of highly expressed genes (e.g., rRNAs, cpDNAs) allowed for the sequencing of as many different genes as possible and those that may be expressed at low levels. However, because of this normalization no inferences of gene expression levels could be made

using just the transcriptome data alone. These datasets have provided the reference for not only this and future RNA-seq experiments, but also the basis for full genome assemblies for both *Caa* (nearly complete) and *Cab* (complete).

To begin these experiments, both common-garden and reciprocal transplant experiments on natural granite and serpentine soils were performed using *Caa* and *Cab* (fig. 12). RNA-seq analyses were conducted to calculate global expression patterns and identify differentially regulated genes that may potentially play a role in serpentine adaptation. Eighteen independent analyses were assembled and annotated. For this dissertation, priority was given to the three most obvious questions: 1) which genes are constitutively expressed in CAB, 2) which genes are induced in CAB on serpentine outcrops, and 3) which genes are induced in CAA on serpentine outcrops? Genes constitutively expressed in CAB in all growth conditions are hypothesized to be essential to survival in serpentine soil. Any 'cost' associated with such expression outweighs the potential of these genes not being turned on. Genes that are induced in CAB on serpentine soils result directly from the environmental conditions. Lastly, genes induced in CAA on serpentine most likely play a role in stress response allowing this species to survive for a short period of time, but not enough of a response to confer full tolerance to these outcrops.

RNA-sequencing (RNA-seq) has revolutionized molecular biology and how transcriptome data are analyzed. This powerful tool allows for the quantification of transcript abundance (or lack thereof) and the detection of differentially expressed genes in taxa or in a variety of environmental conditions (Brown, et al. 2017). RNA-seq offers an advantage over other gene expression methods as a reference is not required and reads can be mapped *de novo*. This is ideal for non-model organisms which most likely lack an annotated genome or transcriptome (Wang, et al. 2009). CAB is not yet a model organism but one major goal of the

Pepper lab is to introduce it as the system for understanding adaptation to serpentine soils. Notably, reference transcriptomes and genomes are available for both sister taxa. Currently, both reference transcriptomes (Hawkins et al. 2017) and reference genomes for both taxa (Garza et al in prep) are available for use. The reference genomes produced a comprehensive set (44,066 genes) of full length coding sequences which was used as the reference for these RNA-seq experiments.

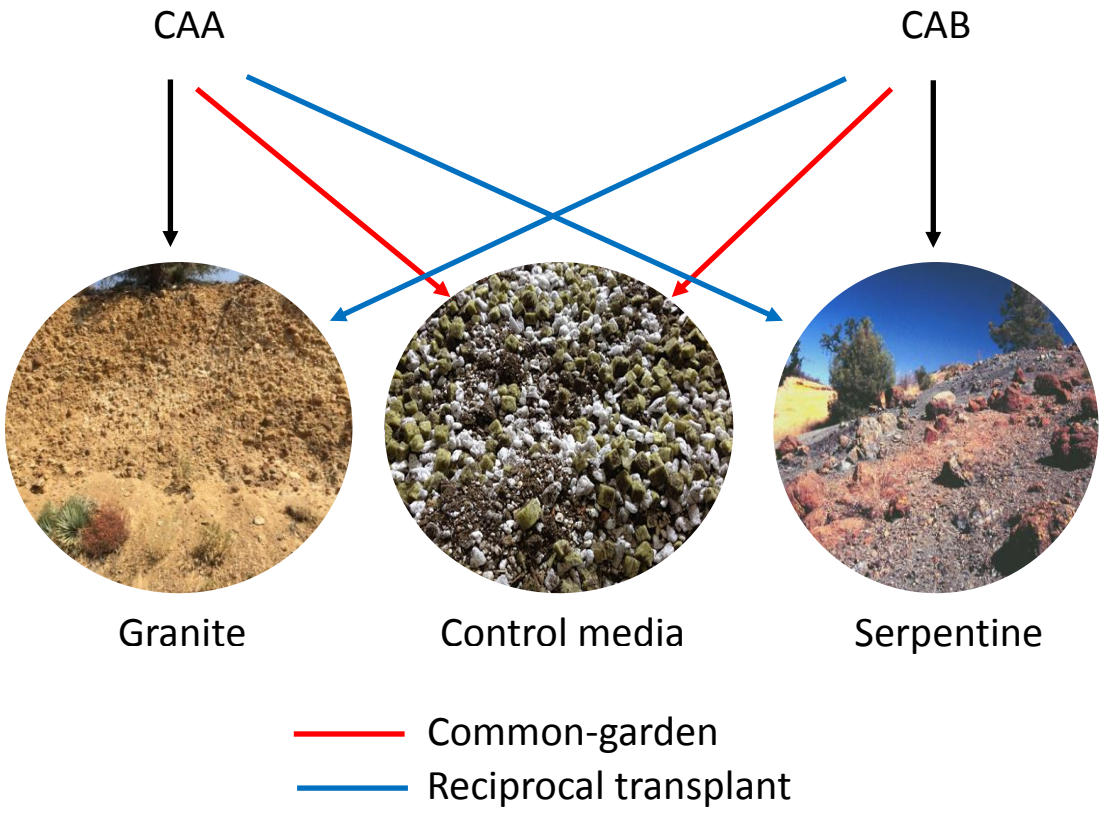

Figure 12. Design for common-garden and reciprocal growth experiments. Three biological replicates for both CAA and CAB were grown in control media (1/4x MS media) as a common-garden experiment and in non-native soils for reciprocal transplant experiments, and in native soils (black lines).

Sequencing resulted in a massive amount of data: 3 biological replicates of both genotypes on each of 3 different environments for 18 independent libraries. An F1 was included in these experiments to gain allele specific data; however, this required a more developed CAA reference genome than what was available at the time this dissertation was written. Therefore, the F1 was excluded from analyses and will be included in future endeavors. This work will show results from just CAA and CAB and answer the three questions above. RNA-seq reads were mapped to a reference set of transcripts built from the CAB genome and gene expression levels were calculated as transcripts per million (TPM) by the program Salmon. EdgeR was then used to calculate log fold change for all genes. Weighted correlation network analysis (WGCNA) was performed to cluster all genes based on their expression profile. These modules were further processed using fisher's exact test to find any overrepresented GO terms.

Materials and Methods

Plant material and growth conditions

Granitic soil for plant growth experiments was collected from the root zones of three CAA plants growing at the original source location of the CAA1 parental inbred line (Los Angeles County, CA, USA; (Pepper and Norwood 2001). Similarly, serpentine soil was collected from the root zones of three CAB plants located near original source of the CAB1 parental inbred line (Santa Barbara County, CA, USA (Pepper and Norwood 2001). For each soil type, samples were combined and mixed to form composite granite and serpentine soils, and then loaded into RLC4U Ray Leach containers (Stuewe and Sons, Tangent, OR, USA) that were pre-loaded with a 7 cm plug of washed and sterile course perlite. Initial hydration of natural soils and all

subsequent watering were carried out with 0.2 μm filter-sterilized Milli-Q ultrapure dH_20 (<18 MΩ-cm; EMD Millipore Corp., Billerica, MA, USA).

To provide control 'replete' growth conditions, plants of each genotype were grown in an artificial culture composed of washed, high-silica sand irrigated with ¼ strength Murashige and Skoog (MS reference) basal salts, macro- and micronutrients, pH 5.8 (Burrell and Pepper, 2012). For all experiments (table 4), seeds were germinated on the surface of the solid medium, in contact with a ~1cm^3 cube of rockwool (Grow-Cubes, Grodan) that had been extensively washed with dH_20. Plants were grown under 8 hours day length in a temperature regime of 22°C (day)/18°C (night). Plants were hydrated to saturation with either diH20 (natural soils) or 1/4x MS media (sand control) every 48 hours, and 4 hours prior to each tissue harvest.

Table 4. Matrix of soil conditions for growth experiments

Genotype	Condition	Reps
CAA	Granite	3
CAA	Serpentine	3
CAA	Media	3
CAB	Granite	3
CAB	Serpentine	3
CAB	Media	3

RNA isolation and sequencing

Shoot and root tissue from seedlings were harvested by flash freezing in liquid N2 at the time of first visible emergence (~1mm) of the 7^{th} true leaf. CAA plants on serpentine soil did not form a 7th true leaf, and were harvested at the time age-matched CAB plants formed the 7th true leaf. All harvests were performed at subjective noon. Root samples were washed briefly in a stream of dH20, then patted dry (using Kim-wipe) prior to freezing in liquid N2.

RNAs were isolated using the RNAqueous kit with plant RNA isolation aid (Thermo Fisher Scientific). DNAs were removed using an on-column DNAse digestion (PureLink DNAse, Thermo Fisher Scientific). RNA quality was examined using the Agilent Tapestation (Agilent). Samples with a minimum RIN equivalent (RINe) score of 7.0 were used for library preparation. RNAseq libraries were prepared using the KAPA Stranded mRNA Kit for Illumina. Library sequencing was performed by the Hudson-Alpha Institute for Biotechnology using the Hi-Seq 2500 instrument.

Data processing (trimming, quality control, RNA quantification)

RNA-seq data were first subjected to quality filtering using trimmomatic (Bolger, et al. 2014) with Illuminaclip active, and a sliding window quality trimming with a window size of 4 bases and an average quality across the window of Q25. Minimum quality to keep both leading and trailing bases was Q15; minimum length (MINLEN) was set to 50 bases.

Gene expression levels were calculated as transcripts per million (TPM) by using the default settings of the program Salmon 0.7.2 (Patro, et al. 2017) on Galaxy (Afgan, et al. 2016). Salmon quantifies the expression of genes using a reference set of transcripts and RNA-seq reads. One advantage of using Salmon for quantification is that the program uses quasi-mapping which makes the process extremely fast as reads are mapped without a base to base alignment (Srivastava, et al. 2016). Another advantage is that quantification of expression levels are computed in transcripts per million (TPM). TPM has become the preferred RNA quantification estimator (rather than RPKM or TPKM) as it normalizes for gene length first, then sequencing depth allowing for the sum of TPMs in each sample to be the same and independent of the mean expressed transcript length (Wagner, et al. 2012).

TPM data for all replicates in all environments were indexed into one file, rounded to the nearest whole number, and then used to run edgeR 3.14.0 (Robinson, et al. 2010) with default parameters, also on the Galaxy platform. edgeR is one of the most widely used programs to calculate differential expression patterns of the RNA-seq data. One major advantage of this program is that it handles small data sets (fewer number of samples) with ease and it can detect transcript specific variation which is important as one outlier among replications can distort interpretation of expression levels (Chen, et al. 2007).

Transcripts in all three environments were filtered based on FDR (<0.01) and $\log_2$ -2 - through +2 (4x change) in expression level. These genes (5,836 total) were incorporated into a final analysis using weighted correlation network analysis (WGCNA) (Zhang and Horvath, 2005). This program uses expression level patterns to cluster genes into modules based off of their similarity. WGCNA was implemented using default parameters in R. First, replicates are examined for any outliers based on their TPM value obtained from Salmon; then, each transcript is placed into one of several modules that best fit the data based on patterns of gene expression. Modules were subjected to Fisher's exact test in Blast2Go to detect any GO terms that may be overrepresented in one cluster compared to all others. Transcripts with similar expression patterns and overlapping GO terms may identify potential pathways to serpentine adaptation.

Results and Discussion

Quantification of gene expression levels using Salmon and edgeR

Eighteen independent Salmon runs were performed to map RNA-seq reads to the reference set of transcripts built from the CAB1 genome assembly and TPM values were calculated for all replicates, of all genotypes, and all environments. Salmon output data was input into edgeR and

differential expression data were calculated as log fold change (logfc). Initially, edgeR generated a multi-dimensional scaling (MDS) plot (fig. 13), which plots the differential expression patterns of separate samples. Replicates of the same sample should show similar expression patterns and cluster together confirming that the difference between groups is greater than within groups.

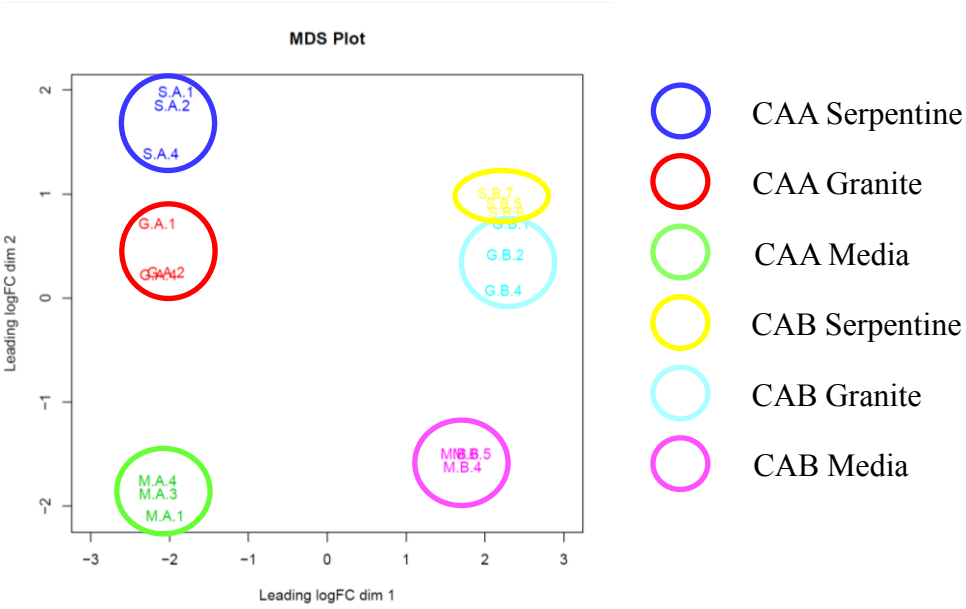

Figure 13. MDS plot of edgeR data. Replicates of the same sample cluster together indicating that there is more variation between samples than within. All CAA samples are on the left and all CAB (right). CAA and CAB are separated by 4 units (x-axis) which equates to a 16-fold leading fold change (2^4=16) between the sister taxa.

Genes found to be the most differentially expressed, filtered based on FDR (<0.01) and log -2 - +2 (4x change), from all environments were then subjected to further analyses using WGCNA.

Gene clustering using WGCNA

Weighted correlation network analysis (WGCNA) was used to cluster genes into modules based on their expression profile. Prior to the analysis, replicates were screened for any outliers that

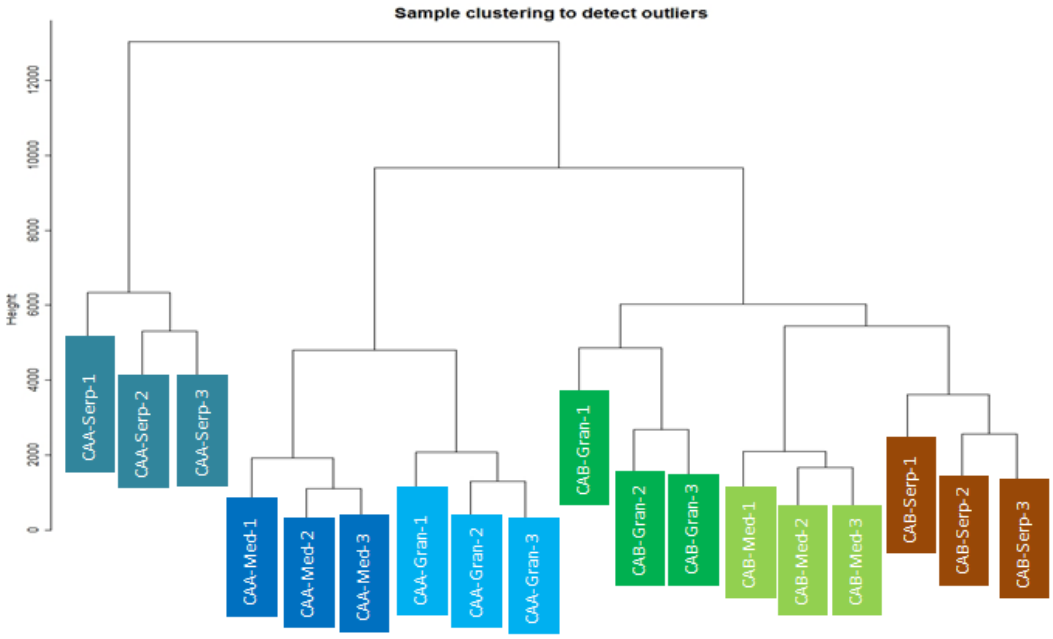

Figure 14. WGCNA clustering to detect outliers. No outliers were detected prior to running WGCNA. All replicates clustered together as well as CAA to other CAA and CAB to other CAB. This shows that all replicates show similar expression patterns and results should not be skewed as no outliers were detected.

might affect the clustering and none were detected as all replicates grouped together (fig. 14). The genes clustered into 18 modules based off of similar expression patterns (table 5a). Each module includes an eigengene which best represents that cluster (table 5b). These genes are informative; however, caution should be taken to avoid basing the entire module from one gene as much variation exists in the functionality of genes in each cluster. These data were critical in answering the three main questions of this chapter.

Table 5a. Expression patterns for modules. The 18 clusters and expression patterns for each cluster. Green indicates genes that are downregulated while red means upregulated. Superscript definitions are as follows: 1 are genes constitutively expressed in CAB, 2 are genes induced in CAB on serpentine, and 3 are genes induced in CAA on serpentine.

Module Name	CAA Media	CAA Serp	CAA Gran	CAB Media	CAB Serp	CAB Gran	Expression Pattern
Black							Constitutively up CAA, low to no CAB
[1]Blue							Constitutively up CAB, low to no CAA
[2]Cyan							Up CAB serp/media, down CAA
Dark green							Up CAA Gran, Media, down Serp; Not in CAB
[2]Dark red							Up CAB S and both A and B Gran
[3]Green							Constitutively up CAA, low to no CAB
Green yellow							Low across board
[3]Grey 60							Up CAA Serp
[3]Light cyan							Very up CAA esp. Serp
Light yellow							Up CAA gran, No to low CAB
[1]Magenta							Constitutively very up CAB
Midnight blue							Up CAB gran, low to no elsewhere
[1]Purple							Constitutively up CAB, very high media; low CAA
Red							Constitutively very up CAA, low to no CAB
Tan							Up CAA media/Gran; no CAB expression
[3]Royal blue							Very up in CAA on serp, low CAB gran
[1]Turquoise							Very up in CAB Constitutively

5b. Eigengenes and annotations for modules.

Module Name	Num Genes	Eigengene	TAIR_ID	Annotation
Black	328	CAB_00020249	AT1G48600.1	PMEAMT, AtPMEAMT | S-adenosyl-L-methionine-dependent methyltransferases
[1]Blue	1251	CAB_00041267	AT4G35785.2	RNA-binding (RRM/RBD/RNP motifs) family protein
[2]Cyan	141	CAB_00002308	AT4G09460.1	AtMYB6, MYB6 | myb domain protein 6
Dark green	61	CAB_00004897	AT5G58784.1	Undecaprenyl pyrophosphate synthetase family protein
[2]Dark red	66	CAB_00017097	AT2G28990.1	Leucine-rich repeat protein kinase family protein
[3]Green	484	CAB_00026520	AT2G40270.1	Protein kinase family protein
Green yellow	201	CAB_00008406	AT2G25600.1	SPIK, AKT6 | Shaker pollen inward K+ channel
[3]Grey 60	578	CAB_00040045	AT1G14780.1	MAC/Perforin domain-containing protein
[3]Light cyan	128	CAB_00037973	AT3G17790.1	ATACP5, ATPAP17, PAP17 | purple acid phosphatase 17
Light yellow	70	CAB_00039603	AT1G13750.1	Purple acid phosphatases superfamily protein
[1]Magenta	298	CAB_00024770	AT5G04490.1	VTE5 | vitamin E pathway gene 5
Midnight blue	137	CAB_00025707	AT5G41780.1	myosin heavy chain-related
[1]Purple	253	CAB_00032850	AT5G18060.1	SAUR-like auxin-responsive protein family
Red	847	CAB_00037786	AT2G29630.3	THIC | thiaminC
Tan	176	CAB_00035147	AT3G46780.1	PTAC16 | plastid transcriptionally active 16
[3]Royal blue	69	CAB_00039818	AT1G11220.1	Protein of unknown function
[1]Turquoise	747	CAB_00019856	AT5G43850.1	ATARD4, ARD4 | RmlC-like cupins superfamily protein

Question 1: Which genes are constitutively expressed in CAB?

WGCNA analyses resulted in 4 modules (blue, magenta, purple, and turquoise) with similar expression patterns: constitutive expression in CAB in all environments. Eigengenes for these modules were: "AT4G35785.2" RNA-binding (RRM/RBD/RNP motifs); "AT5G04490.1" VTE5 vitamin E pathway gene 5 family protein; "AT5G18060.1" SAUR-like auxin-responsive protein family; and "AT5G43850.1" ATARD4, ARD4 RmlC-like cupins superfamily protein, respectively. Fisher's Exact Test was used to determine if any GO terms were overrepresented in these genes. There were several terms that were overrepresented (fig. 15) and, surprisingly, most of them are associated with chloroplast and plastid functionality.

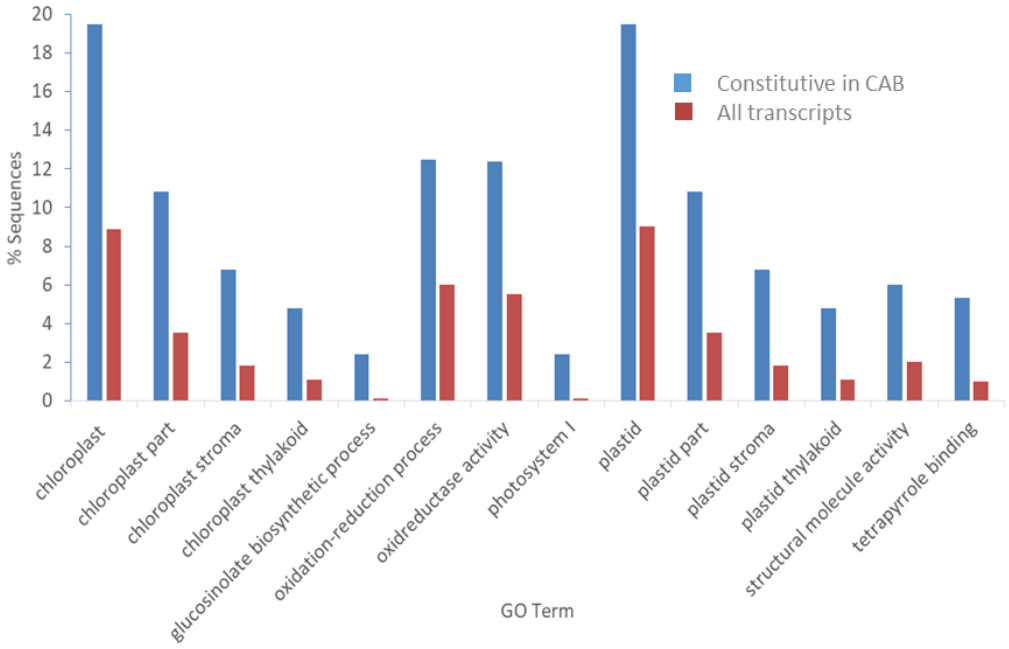

Figure 15. GO terms overrepresented in constitutively expressed CAB genes. Most of these genes have something to do with photosynthesis (FDR <0.05).

The enrichment of photosynthetic related GO terms in constitutively expressed genes in CAB on serpentine is not only unexpected but actually counterintuitive. Environmental stress

usually leads to the downregulation of photosynthesis. Unfavorable environmental conditions, such as those found in serpentine soils (e.g., drought, excess heat, heavy metals, and high amounts of light) causes damage to photosynthetic machinery and decreases in productivity (Gururani, et al. 2015). Further, biotic stresses such as insects, fungal infections, and viruses leads to global downregulation of photosynthetic genes (Bilgin, et al. 2010). With all of the environmental stresses that CAB plants have adapted to, this was a confounding result. However, a complex series of interactions may offer an explanation.

Serpentine soils are notoriously barren as they have little vegetative coverage (see fig. 1). This dearth of coverage exposes CAB plants to extremely high levels of sunlight. Exposure to high levels of light prompts the production and accumulation of anthocyanins (a class of flavonoid pigments giving plants coloration) in vegetative tissues (e.g., leaves) as a defense mechanism to protect against possible photo-inhibition (Das, et al. 2011). A distinct morphological characteristic unique to CAB plants is their variegated and mottled leaves rich with anthocyanins (fig. 16), which most likely help the plants deal with the high light environment. But there is always a trade-off and here it happens to be the expense at which anthocyanins are made as it requires a lot of energy which it obtains carbon assimilation (Drumm-Herrel and Mohr 1985). Constitutive expression of photosynthetic genes in CAB in all environments could be a 'bet-hedging' technique to ensure the plant produces the anthocyanins for protection while still making enough energy for other functions. It has long be assumed that this accumulation of anthocyanins seen in CAB plants was due to the lack of phosphorous in serpentine soils, as this is a direct cause of this phenotype (Jiang, et al. 2007). However, exposure to this much sunlight may offer a different, or at least additional, explanation.

Figure 16. CAB plant on natural serpentine outcrop. Anthocyanin accumulation is seen in a mottled pattern in leaves. Photo courtesy of and with permission granted from John Moule, 2017.

Ethylene is a natural plant hormone with several important functions (e.g., fruit ripening, nutrient cycling, and responding to a multitude of environmental stresses). Additionally, four different genes related to ethylene production and/or movement show constitutive expression patterns in CAB: ethylene responsive element binding factors 1, 8, and 15 (ERF1, ERF8, ERF15) and ethylene-forming enzyme (EFE). Again, this is a confounding result as the ethylene usually decreases expression of photosynthetic genes to help deal with the reactive oxygen species (ROS) that result from carbon assimilation. However, these genes most likely have pleiotropic effects. ERF1 has been shown to function in both abiotic (Cheng, et al. 2017) and biotic stress responses (Berrocal-Lobo, et al. 2002). Cheng et al. (2017) found that 35S:ERF1 transgenic *Arabidopsis* plants were better able to deal with several environmental stresses including drought, salt, and heat, all of which CAB deals with. Further, high salinity and drought

treatments induced the rapid and transient expression of ERF1. ERF8 has been shown to play a vital role in leaf senescence (Koyama, et al. 2013) which is the final stage of leaf development where nutrients are moved from old to young leaves. Leaf senescence is highly dependent on age, but can also result from environmental stresses, such as those seen in serpentine soils. Constitutive expression of this gene may help CAB plants mobilize any nutrients they acquire and shuffle to newly developing leaves. ERF15 is a positive regulator of the abscisic acid (ABA) response (Lee, et al. 2015). During normal plant growth, ABA expression is usually quite low but it does control seed development. When plants are stressed (e.g., drought, high salinity, and excessive heat), the level of expression increases significantly (Lee, et al. 2015). Again, CAB plants are exposed to all three abiotic stresses and thus constitutive expression of this gene to upregulate ABA response makes sense. Finally, EFE is a very important gene responsible for the biosynthesis of ethylene. EFE is upregulated in environments with limited macronutrients such as nitrogen, potassium, phosphorus, and magnesium as well as those with high levels of metals such as copper, cadmium, and selenium (De Gernier, et al. 2016). Constitutive expression of EFE in CAB makes perfect sense and should be considered a high priority candidate gene that confers tolerance to serpentine soil.

Question 2: Which genes are induced in CAB on serpentine?

Salmon analyses implicated only a small set of genes (245 total) that are upregulated in CAB on serpentine when compared with common-garden media (fig. 17; table 6). This indicates that these genes are induced in serpentine soils most likely to deal with environmental stresses. This expression pattern was found in the cyan and dark red modules with eigengenes “AT4G09460.1”

AtMYB6, MYB6 myb domain protein 6 and "AT2G28990.1" Leucine-rich repeat protein kinase family protein, respectively.

The MYB gene family, one of the largest found in plants, has well over 100 genes and includes functions from cell growth to cell death (Stracke, et al. 2001). MYB6 is a transcriptional repressor of the Production of Anthocyanin Pigment 1 (PAP1) gene. PAP1 is a well-established gene that functions as a positive regulator of anthocyanin production (Sharma and Dixon 2005). The most obvious phenotypic difference between CAA and CAB is sepal color (see fig. 1) as the serpentine endemic lacks the deep purple color found in its sister taxa. The discovery of induced upregulation of MYB6 is not surprising, but whether sepal color or anthocyanin production or repression is an adaptive trait for serpentine plants remains unknown.

Leucine–rich repeat (LRR) proteins, another large family of plant genes, have long been known to function, in a broad aspect of plant defenses (Jones and Jones 1997). Not much is known about this specific LRR. However, an earlier study by (Shahollari, et al. 2007), investigated the relationship between *Arabidopsis thaliana* and an endophytic fungus, *Piriformospora indica.* The authors concluded that co-cultivation of seedlings of *A. thaliana* along with *P. indica* resulted in increased fresh weight of both the roots and shoots. In addition to this increase in growth, a massive intake of phosphate from growth conditions was observed. During this period, several genes were transiently induced, one of them being this particular LRR. However, taxa in the Brassicaceae family are assumed to be non-mycorrhizal as no species has been naturally found to participate in the mutualistic relationship (Regvar, et al. 2003). The induction of this LRR in CAB in serpentine may assist in seed establishment, growth, and nutrient acquisition. Future studies should address this as this could be an interesting candidate gene for adaptation to serpentine.

Additionally, a suite of genes upregulated in CAB on serpentine was completely missing in CAA grown on serpentine soils (table 6). These genes may be the most important for adaptation to serpentine soil. Interestingly, some of these genes negatively regulate some defense response genes (only in CAB and only on serpentine). This result seems odd/counterintuitive as defense genes would most likely be needed for survival; however, these particular defense genes perhaps are not required for tolerance and are downregulated to save energy for genes that are needed. Furthermore, more than 50% of these genes have no hit to anything available in published databases. This could indicate orphan genes unique to CAB but additional experiments need to be performed to make this statement with confidence.

Finally, CAA and CAB share a similar number of upregulated genes (216) in serpentine conditions (fig. 17, table 8). Two of these genes are members of the important MYB family (MYB51 and MYB59). They are transcription factors that respond to a number of different environmental stresses including the acquisition of nutrients and the intake and cycling of heavy metals. The presence of these genes in CAA grown on serpentine soils is most likely the only reason the seeds can germinate and grow for a few days before dying. CAA plants grown in serpentine conditions have a 100% mortality rate and have never transitioned from vegetative to reproductive phases, and therefore have never produced flowers or offspring. They appear to be an evolutionarily dead end. Genes upregulated in serpentine vs media and shared between CAA and CAB may be related to ancestral stress responses obtained from their common ancestor.

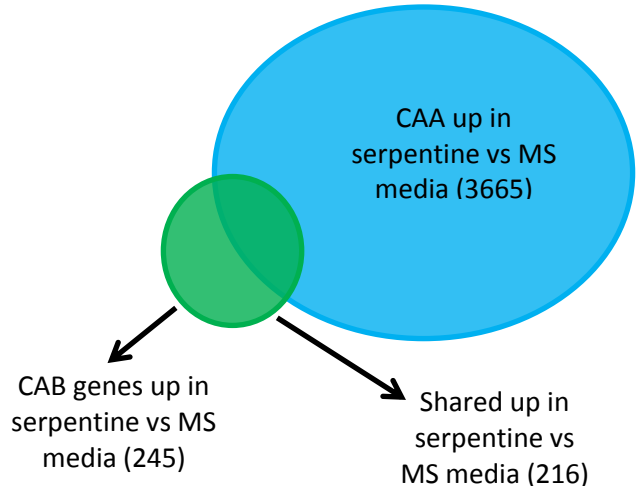

Figure 17. Venn diagram of differentially expressed genes in CAA and CAB. Circles show number of genes induced in only in CAA on serpentine, only in CAB on serpentine, and those induced and shared between the sister taxa.

Table 6: Genes upregulated in CAB on serpentine relative to media. Fold change is included. N/A denotes that these genes are completely absent in CAB on media

Fold change	Gene description
N/A	K ion import, responses to cold, salt stress, and drought
N/A	Bifunctional inhibitor/lipid-transfer protein
174X	TPS10 Trehalose phosphate synthase
131X	(TPR)-like: response to oxidative stress
106X	ZIP3 Zinc transporter 3
30X	Heavy metal transport/detoxification superfamily protein
24X	Protein phosphatase 2C family protein
22X	ATNRT2:1 Nitrate transporter
16X	CaLB domain: metal ion binding, phospholipid binding
15X	Phosphate transport, phosphate ion homeostasis
15X	SRPK4: response to oxidative stress
14X	oxidation-reduction process, response to red or far red light
14X	NRT2.6 High affinity nitrate transporter 2.6
12X	Ankyrin repeat family protein
12X	TGA1A-related gene 3: defense response
11X	Cold regulated gene 27: response to cold, response to karrikin
9X	ATMYB90, PAP2, MYB90: production of anthocyanin pigment
4X	CSD2, CZSOD2 Copper/zinc superoxide dismutase 2
4X	WR3 Nitrate transmembrane transporters

Table 7. Genes that are absent in CAA and upregulated in CAB on serpentine. Fold change is included. N/A denotes genes that are only found in CAB on serpentine and not in CAB in media.

Fold change	Gene description
N/A	Alpha/beta-Hydrolases superfamily protein, catalytic activity
N/A	Negative regulation of defense response and protein catabolic process, plant-type hypersensitive response, protein ubiquitination
N/A	ATP-citrate lyase A-3 acetyl-CoA biosynthetic process
N/A	No known function
N/A	No known function
330X	RNA-binding (RRM/RBD/RNP motifs) family protein
305X	No known function
300X	No known function
275X	No known function
75X	Heme and iron binding, glucosinolate biosynthetic process, oxidation-reduction process, response to UV, response to insect

Table 8. Shared genes that are induced only in serpentine conditions.

Fold change	Genes description
137X	Tetratricopeptide repeat (TPR): response to stress
13X	Senescence-associated family protein
9X	AMT2;1, AMT2: ammonium transport
9X	COR413-PM1: response to water deprivation, cold acclimation
9X	Heat acclimation
8X	AGAMOUS-like 21: positive regulation of transcription
8X	MYB59 responses to: cadmium ion, chitin, gibberellin, jasmonic acid, salicylic acid
7X	Zinc/iron-chelating domain protein
7X	NAC domain containing protein: response to jasmonic acid, transcription
7X	Choline transporter family protein: cell to cell communication
7X	Responses to abscisic acid, oxidative stress, salt stress, water deprivation
7X	PRXR1: metal ion binding, hydrogen peroxide catabolic process
7X	MYB51 responses to: defense, abscisic acid, auxin, ethylene, gibberellin, insect, jasmonic acid, salicylic acid, salt stress
7X	AMT1;1: ammonium transport, response to abscisic acid

Question 3: Which genes are induced in CAA on serpentine?

The final question asked for this dissertation is which genes were induced in CAA on serpentine soils. Not surprisingly, a large number of genes (3665 total) were upregulated in CAA

on serpentine soils (fig. 17) and this expression pattern was found in 4 modules: green, grey 60, light cyan, and royal blue. The eigengenes for these modules were: “AT1G14780.1” MAC/Perforin domain-containing protein, “AT3G17790.1” ATACP5, ATPAP17, PAP17 purple acid phosphatase 17, “AT2G40270.1” Protein kinase family protein, and “AT1G11220.1” Protein of unknown function, respectively. Fisher’s exact test implicated many GO terms that were overrepresented (fig. 18) and most of these terms suggest these genes are related to serpentine tolerance. These genes are absent or expressed at very low levels when the same plants are grown in granite or media. This is highly indicative of environmentally induced genes.

Interestingly, several of the same genes that are induced in CAB on serpentine are also induced in CAA in serpentine (table 8). Again, this is indicative of homology from a common ancestor. Access to these genes allows CAA to at least germinate on serpentine, but never mature to an adult plant capable of producing offspring. Among the list of genes induced in CAA on serpentine are a suite of MYB genes (14 total). As stated in the previous section, these genes are important for a number of functions, especially aiding in stress response and anthocyanin production.

Another interesting result is the overrepresentation of GO categories “galactolipid biosynthetic”, “galatose binding”, and “galactolipid metabolic process”. Even in ‘normal’ soils, the amount of phosphate is usually limited and this is obviously exasperated in serpentine outcrops. It has long been known that when phosphate is limited or lacking, plants will increase their production of galactolipids to reduce their need for phosphate (Khozin-Goldberg and Cohen 2006). In phosphate starved *Arabidopsis thaliana*, the galactolipid DGDG is accumulated to accommodate for the lack of the essential plant nutrient (Essigmann, et al. 1998).

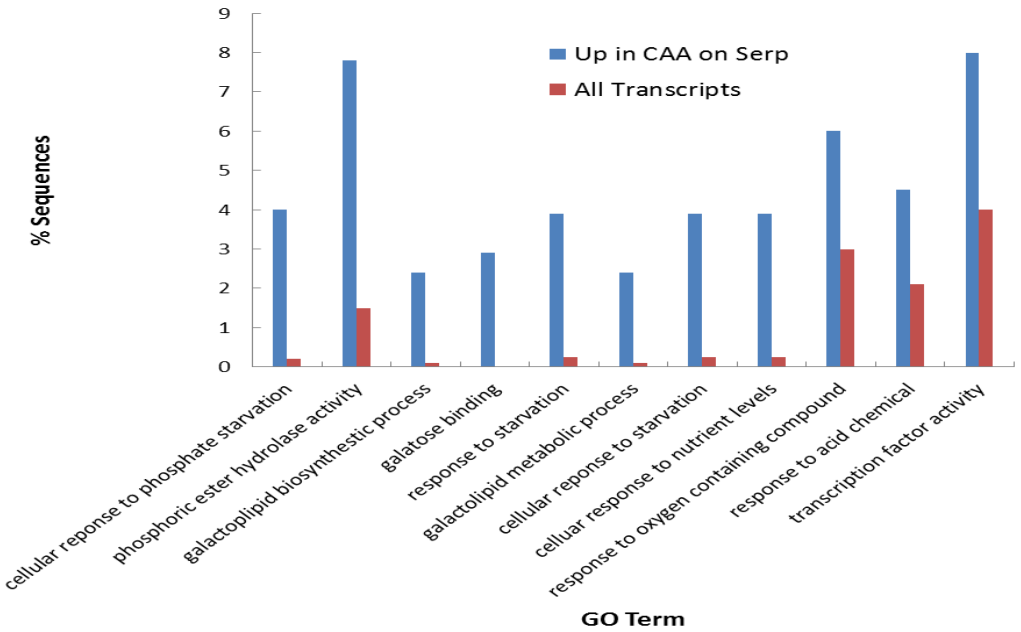

Figure 18. GO terms overrepresented in CAA when induced in serpentine.

The induction of these genes offers a look into the timeline of serpentine tolerance. The lack of essential nutrients is most likely the first set of stressors plants must deal with on serpentine soils. The over expression of genes producing galactolipids offers some sort of protection against phosphate starvation, even in plants not adapted to serpentine soils.

Summary and Perspectives

Reciprocal transplant and common-garden growth experiments, along with a comprehensive RNA-seq analysis, allowed for the quantification of expression levels differences between two sister taxa who are ecologically, morphologically, and geographically differentiated. There exists a wealth of data and questions that can be asked and answered. For the scope of this dissertation, the three most obvious and glaring questions were addressed. First, what genes are constitutively expressed in CAB plants? These genes show high expression

regardless of which environment they grown in. Such investment in always expressing these genes may be indicative of their extreme importance to tolerance of serpentine outcrops.

Table 9. Genes upregulated in CAA on serpentine relative to media. Fold change is included. N/A denotes genes not found in CAA on media or granite.

Fold change	Gene description
N/A	Iron-sulfur binding
N/A	PII-like copper ion binding, response to metal ion
N/A	NIK2 defense response
N/A	ATPAP24: defense response
N/A	Metal ion binding, protein serine/threonine phosphatase activity
N/A	Hop2 stress-inducible protein: response to cadmium ion
N/A	BGAL5: carbohydrate metabolic process
N/A	ATNST-KT1: carbohydrate/sugar transport
>1000X	Tetratricopeptide repeat (TPR)-like superfamily protein
367X	F-box family protein: chloroplast organization
272X	Jasmonic acid mediated signaling pathway, response to water deprivation
114X	EMB93: response to singlet oxygen, chloroplast organization
107X	PDE225, PTAC7: plastid transcriptionally active7
75X	PYR1-like 11: Encodes a member of the pyrabactin resistance
73X	Ubiquitin-like protein: response to nitrogen starvation
68X	MAPK kinase: responses to cadmium ion, osmotic stress, salt stress, wounding
57X	SUS3: response to mannitol, response to water deprivation
57X	Cadmium ion homeostasis, detoxification of cadmium ion, metal ion transport
34X	CIPK9: responses to potassium starvation, cold, mannitol, salt stress, wounding
33X	SART-1 family: cotyledon vascular tissue pattern formation, flower
30X	FPF1-like protein 1: encodes a similar to flower promoting factor 1
30X	Chlorophyll catabolism

Unexpectedly, most of these genes are related to chloroplasts and other plastids. This may result from CAB plants being exposed to high levels of sunlight and increased production and accumulation of anthocyanin as a "sun screen" to protect the plants from photo-inhibition. Genes that were induced from environmental stressors were similar in both CAA and CAB, which was not unexpected. Most of these genes deal with nutrient deficiency or heavy metal tolerance or

transport. CAA had many more genes that were induced, and this was also expected as it was extremely stressed and trying, to no avail, to survive. Further, CAA showed increased expression of a suite of genes related to the production of galactolipids. It has been shown that in the absence or lack of phosphate, a major limiting nutrient in serpentine soils that plants will upregulate their production of galactolipids to accommodate for the lack of phosphate. These data have provided a wealth of insight into selecting high quality candidate genes for further evaluation via transformation/knock out experiments (see future endeavors). Also, inclusion of F1 data will allow for allele specific heredity information and even more in depth knowledge of the adaption to serpentine soils.

CHAPTER IV

CONCLUSIONS AND FUTURE DIRECTIONS

Life on serpentine soils requires a complex suite of adaptive traits in response to several harsh environmental challenges. Dealing with just one harsh condition elicits a genetic response for survival. These plants have to overcome a gauntlet of stresses from low to no essential nutrients, toxic levels of heavy metals, limited moisture, high light levels, and elevated soil temperatures. In addition, no transitions zones exist between serpentine and non-serpentine outcrops. This tolerance has to be quick and absolute for survival. This leads to the hypothesis that many of the genes responsible for survival function pleiotropically. This small plant has the ability to hyperaccumulate and sequester heavy metals from the surrounding soils and be used in phytoremediation efforts. Additionally, the genes responsible for serpentine tolerance, once identified can be introduced into crop plants for more efficient, and much less, fertilizer usage.

Not many organisms can achieve this adaptive regime; however, the few that do usually thrive as they most likely have been outcompeted elsewhere. The genetic basis for this adaptation is largely unknown and it is a top priority in the Pepper lab is to establish CAB as the model organism for understanding tolerance to extreme environments. This serpentine endemic is a close relative of *Arabidopsis thaliana*, a member of one of the most studied and economically important plant families (Brassicaceae), has a small, diploid genome, a short generation time, is interfertile in lab (which allows for the creation of highly informative RILs), is self-fertile, and due to floral structure is amenable to transformation by floral dip and tissue culture.

This dissertation offers a couple of pieces to the puzzle and work by Pepper lab members (current and former) will be used together to enlist a strong set of candidate genes. To date, comprehensive genome sequences for both CAA and CAB have been assembled and annotated by Garza et al. (in prep). Population genomic analyses have been performed by Burrell et al. (2011, 2012, and in prep). All of this data will be incorporated in downstream analyses (see future directions) to get to the root of serpentine adaptation and elicit a set of candidate genes conferring tolerance.

Concluding remarks from chapter II, transcriptome data

Comprehensive reference transcriptomes were built using RNA from CAA and CAB tissues, at various stages of plant development, and under differing environmental conditions (supplemental table 1, in Appendix A). Next generation sequencing technology was employed to perform a *de novo* assembly and annotation of each transcriptome. In all, 93,647 and 83,484 RTLs resulted for CAA and CAB, respectively. These RTL were then further subjected to reciprocal best blast hit (RBH) to find orthologs between CAA and CAB. RBH implicated approximately 29,500 tentative orthologous pairs (TOPs) that were used in most subsequent analyses. These analyses resulted in several major conclusions:

1. Comparison of TOPs shows elevated d*N*/d*S* ratios with a global mean of 0.3458. This mean is unusually high for sister taxa that have diverged recently (see table 1). High d*N*/d*S* ratios are usually indicative of positive selection; however, several other mechanisms must be considered. Genetic drift (small population sizes) can overwhelm purifying (negative) selection in otherwise conserved sequences. When the same set of genes was compared in CAA and CAB and in *A.*

lyrata and *C. grandiflora*, CAA/CAB had higher d*N*/d*S* across the board. This phenomenon occurred across a broad set of genes which helps support that they have higher *dN/dS* due to reduced efficiency of purifying selection due to drift.

Additionally, relaxed negative selection, along with the accumulation of slightly deleterious alleles on genes no longer needed may inflate *dN/dS*. Twenty sets of genes with functionality in shade avoidance showed evidence of relaxed negative selection (see table 2). CAB plants grow on barren outcrops with a lot of open vegetation. These plants most likely never have to deal with or worry about shade avoidance. These genes are energetically expense to make and maintain so the accumulation of deleterious alleles could result from nonfunctional genes.

A set of TOPs were found to be under positive selection however (see figure 5). GO enrichment implicates a set of transcription factors and signal transduction molecules that are most likely under positive selection as the result adapting to very different edaphic environments. GO enrichment also suggested purifying selection was working on keeping a set of catabolic and anabolic enzymes free of deleterious mutations. These genes are most likely mportant for the survival of these, and really any, plants regardless of environment.

2. d*S* estimates between orthologs shows evidence of reticulate evolution (see figure 9). Since the rate of synonymous substitution is a proxy for mutation rate, the relative timing of divergence between CAA and CAB was estimated in both chloroplast and nuclear genomes. Chloroplast orthologs had a much more recent divergence event with an adjusted mean of d*S* = 0.0003. Most nuclear genes had d*S* values almost 4 fold higher than chloroplast genes and had a strongly bimodal distribution of d*S*. These two distinctive peaks suggest that after an older divergence

CAA and CAB were reintroduced and introgression occurred. Geographic and climate data suggest that during colder and wetter periods of time, populations may have expanded and allowed CAA, CAB, and even members of serpentine tolerant *Strepanthus* to come into contact and hybridize. Serpentine tolerant genes from *Strepanthus* may have been inherited by CAB via reticulate evolution.

3. d*S* estimates between paralogs indicates two distinct periods of extensive gene duplication (see figure 10). All against all BLAST analyses implicated a set of 1299 and 729 paralogs in CAA and CAB, respectively. Mean d*N*/d*S* estimates between paralogs were significantly lower than orthologs. This suggests that these genes are under negative selection to retain functional copies of both genes. When d*S* values between TPPs were graphed, a bimodal distribution was again observed indicating two distinct duplication events. The older peak occurred during a well-known Brassicaceae whole-genome duplication event. The second peak indicates a much more recent duplication event in both taxa. Both genes in paralog pairs actively being transcribed and evolving via negative selection indicates that each copy is providing a needed function for these plants. Duplicated genes can take on new or different functions or change gene expression levels. Further, it is well known that gene duplication simulates evolutionary innovation, thus, shows the potential importance of these duplications.

Final comments on transcriptome data

Assembling and annotating these comprehensive reference transcriptomes was fundamental in the process of making CAB the model system to explain serpentine tolerance. The data observed here, along with genome sequences, QTL mapping, and RNA-seq experiments will afford a

highly confident list of candidate genes that confer tolerance to serpentine soils. This data was peer reviewed and published in Genome Biology and Evolution in December, 2017.

Concluding remarks from chapter III, RNA-Seq data

Common-garden and reciprocal transplant growth experiments were performed on CAA, CAB, and a F1 hybrid. Seeds of each genotype were grown in natural granite and serpentine soils in additional to MS media. CAA has a 100% mortality rate when grown under serpentine conditions; however, enough tissue and RNA was available for extraction for all analyses. RNA-seq data was mapped to a reference set of transcripts created by the CAB genome assembly. Salmon was used to calculate gene expression levels using TPM method, edgeR was implemented to determine differential expression patterns among genotypes and environments, and WGCNA was use to cluster genes bases on their expression profile. In total, the genes clustered in 18 modules and this data was integral for answering the three questions this dissertation asks: 1) which genes are constitutively expressed in CAB, 2) which genes are induced in CAB on serpentine, and 3) which genes are induced in CAA on serpentine.

1. Which genes are constitutively expressed in CAB?

Four modules had expression profiles that suggested constitutive expression in CAB regardless of environment. Genes in these modules were subjected to Fisher's Exact Test to determine if any GO terms were overrepresented. There were several terms that were overrepresented (see figure 15) and, surprisingly, most of them have some function related to photosynthesis. This pattern of constitutive expression of genes related to photosynthesis may be explained by the lack of coverage on serpentine outcrops and the extremely high light levels. When plants are

exposed to high light they produce and accumulate anthocyanins to act almost like a sun screen and prevent photo-inhibition. CAB plants lack any anthocyanins in floral tissue but are notorious for having mottled patterns of deep purple in their leaves (see figure 16). It was assumed that this accumulation of anthocyanins was a result of phosphate deprivation, and it most likely plays a role, but these new data suggest at least an additional explanation. These anthocyanins are useful and helpful but producing them requires a lot of carbon. CAB plants may have to increase the process of photosynthesis and carbon assimilation to account for this huge carbon sink.

Additionally, four genes related to the production of ethylene were found to be constitutively expressed in CAB plants. Ethylene provides a multitude of functions for plants and is well known to deal with both biotic and abiotic stresses. Upregulation of ethylene can help plants deal with everything from fungal and viral infections to protection from herbivores. Abiotically, ethylene helps plants deal with drought, salt, and heat stress. In addition this hormone functions in leaf senescence and nutrient cycling. CAB plants may use ethylene to transport any possible available nutrients to younger, growing leaves. Finally, EFE shows constitutive expression in CAB. This gene has been shown to be over expressed in plants grown in conditions that emulate serpentine soils (e.g., low essential nutrients and high levels of toxic metals). EFE should be considered a high priority candidate gene that confers serpentine tolerance.

Genes that are under constitutive expression regardless of environment should be considered vastly important to serpentine tolerance. Over expression of these genes may be energetically costly, but ‘normal’ or low/no expression may be lethal.

2. Which genes are induced in CAB on serpentine?

Two modules suggested a small set of genes that are induced in CAB on serpentine. This indicates that these genes are induced in serpentine soils most likely to deal with environmental stresses. The eigengenes for these two modules (MYB6 and LRR) offer interesting results. MYB6 is a transcriptional repressor of the PAP1 gene which produced anthocyanin. This is counterintuitive to the data found in question one as an increase in anthocyanin is seen in CAB leaves. However, CAB lacks anthocyanin in its sepals rendering it to be almost always pure white. MYB6 may repress the production of anthocyanins in sepals to increase production in leaves. As for the leucine rich repeat, this particular one was one of several genes induced when *A. thaliana* was exposed to *P. indica*. This exposure caused *A. thaliana* to increase in size and uptake a massive amount of phosphate. This gene may play an important role in nutrient acquisition and should be considered a candidate gene that confers serpentine tolerance.

Further, some genes induced in CAB on serpentine were completely missing from CAA (see table 7). Interestingly, more than half of these genes have no known function to any published database. This could indicate that these genes are unique to CAB; however, further investigation needs to be done to support this. If these genes are truly unique to CAB, they could be some of the most important genes that confer serpentine tolerance. It would be interesting to see if some of these genes are shared among any unpublished data sets from any other serpentine tolerant or endemic plant species.

Finally, a small number of genes were shared between CAA and CAB in serpentine conditions (see figure 17 and table 8). Most of these genes deal with abiotic stresses including response to nutrient starvation and heavy metal exposure, heat stress, and senescence. These results were expected and since both sister taxa induce these genes in serpentine this may

indicate that they are related to ancestral stress responses that were inherited from a common ancestor.

3. Which genes are induced in CAA on serpentine?

Over 3600 genes were induced in CAA on serpentine, which is not surprising as this plant has a 100% mortality rate when grown in this condition. CAA plants are extremely stressed and try anything possible to survive. Most of these genes in these modules are completely absent in CAA when grown in media or granite conditions. Further, many of these genes are the same or similar to ones induced in CAB on serpentine. Interestingly, several GO terms related to galactose production were overrepresented when compared to other modules. In phosphate starved conditions, plants will increase production of galactolipids to accommodate and provide some sort of help.

The induction of these genes in CAA may result from common and shared stress genes generally found in plants. These genes react to the environmental stresses but are not enough to confer tolerance. They could explain why CAA germinates in serpentine conditions and grows just enough to get viable tissue for RNA extraction.

Final comments on RNA-seq data

This experiment resulted in a massive amount of data. There are dozens of questions that can, and should, be addressed. For brevity, this dissertation focused on three of the major questions surrounding gene expression patterns for these plants and in these environments. WGCNA analyses offered a clear and descriptive way to group and visualize the data. These modules offer

so much detail and will be further explored in subsequent analyses. These experiments are far from complete and will be even more informative upon additional review.

Future directions

1. Continue analyses of all RNA-seq data

An F1 hybrid was included in all growth experiments but was not discussed in the dissertation because at the time of writing reassembly of the CAA genome was still being performed. Both genomes should be included when analyzing the hybrid so allele specific data can be assessed. With both genomes available, inheritance patterns can be interpreted and which parent provided each allele for the hybrid can be determined.

Further, many questions can still be asked about the wealth of data provided by WGCNA. For example, which genes are downregulated when CAA is grown on serpentine soils? This may provide some insight why CAA has complete mortality when grown in these conditions. Are there genes expressed in CAA but not CAB? Which genes are downregulated in CAB? Additionally, genes only found in CAB with no hits to any published database should be examined further. The first step would be to map them to the genome assembly to see if they are actually present and not artifactually transcribed. Then these genes can then be analyzed for coding potential and if found further examined. It may be helpful to blast to other known serpentine plants. If these genes are found in other plants, they may be likely to confer tolerance to serpentine soil.

Finally, phenotypic and genotypic data from growth experiments using RILs is currently being examined to assess QTL underlying nickel and low calcium tolerance. Future experiments will include additional environmental stresses. QTL will be examined to see if any genes within

those regions show differential expression. If so, they may be some of the most biologically important candidate genes to further investigate.

2. Test functionality of candidate genes by transformation into a knock-out *A. thaliana* line. A set of high priority candidate genes found by incorporating data from transcriptome and genomes annotations, population genetic data, QTL mapping, and RNA-seq analyses is being determined. These genes will be transformed in *Arabidopsis thaliana* knock-out lines to determine phenotype. Preliminary transformation experiments have been performed and were successful. Future experiments will use long PCR methods and product will be cloned into *Agrobacterium tumefaciens*, a gram negative soil bacteria, carrying the pCambia3300 vector. Transformed bacteria will be grown in a liquid culture, allowing simple and effective methods for subsequent floral dipping procedures. Transformed lines will be phenotyped side-by-side in varying, stressful environmental conditions.

Validation of candidate genes responsible for serpentine tolerance would provide ground breaking evidence to explain the molecular mechanisms required for adaptation to serpentine soils. Further, they would provide fundamental information needed to attain broader impact goals of phytoremediation and engineering more efficient crop plants. Finally, these genes will allow for CAB to be the model organism for adaptation to stressful environments.

REFERENCES

Afgan E, Baker D, van den Beek M, Blankenberg D, Bouvier D, Cech M, Chilton J, Clements D, Coraor N, Eberhard C, et al. 2016. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. Nucleic Acids Res 44:W3-W10.

Al-Shehbaz IA, Beilstein MA, Kellogg EA. 2006. Systematics and phylogeny of the Brassicaceae (Cruciferae): an overview. Plant Systematics and Evolution 259:89-120.

Alonso JM, Stepanova AN, Leisse TJ, Kim CJ, Chen H, Shinn P, Stevenson DK, Zimmerman J, Barajas P, Cheuk R, et al. 2003. Genome-Wide Insertional Mutagenesis of *Arabidopsis thaliana*. Science 301:653.

Altenhoff AM, Dessimoz C. 2009. Phylogenetic and functional assessment of orthologs inference projects and methods. PLoS Comput Biol 5:e1000262.

Anacker BL. 2014. The nature of serpentine endemism. American Journal of Botany 101:219-224.

Arbiza L, Dopazo J, Dopazo H. 2006. Positive selection, relaxation, and acceleration in the evolution of the human and chimp genome. PLoS Comput Biol 2:e38.

Arnold BJ, Lahner B, DaCosta JM, Weisman CM, Hollister JD, Salt DE, Bomblies K, Yant L. 2016. Borrowed alleles and convergence in serpentine adaptation. Proceedings of the National Academy of Sciences 113:8320-8325.

Bartish IV, Ainouche A, Jia D, Bergstrom D, Chown SL, Winkworth RC, Hennion F. 2012. Phylogeny and colonization history of *Pringlea antiscorbutica* (Brassicaceae), an emblematic endemic from the South Indian Ocean Province. Mol Phylogenet Evol 65:748-756.

Beilstein MA, Nagalingum NS, Clements MD, Manchester SR, Mathews S. 2010. Dated molecular phylogenies indicate a Miocene origin for *Arabidopsis thaliana*. Proceedings of the National Academy of Sciences 107:18724-18728.

Berrocal-Lobo M, Molina A, Solano R. 2002. Constitutive expression of ETHYLENE-RESPONSE-FACTOR1 in *Arabidopsis* confers resistance to several necrotrophic fungi. The Plant Journal 29:23-32.

Bilgin DD, Zavala JA, Zhu J, Clough SJ, Ort DR, DeLucia EH. 2010. Biotic stress globally downregulates photosynthesis genes. Plant Cell Environ 33:1597-1613.

Blanc G, Wolfe KH. 2004. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. Plant Cell 16:1667-1678.

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

Bou-Torrent J, Roig-Villanova I, Galstyan A, Martínez-García JF. 2008. PAR1 and PAR2 integrate shade and hormone transcriptional networks. Plant Signaling & Behavior 3:453-454.

Bowers JE, Chapman BA, Rong J, Paterson AH. 2003. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. nature 422:433-438.

Brady KU, Kruckeberg AR, Bradshaw HD. 2005. Evolutionary Ecology of Plant Adaptation to Serpentine Soils. Annual Review of Ecology, Evolution, and Systematics 36:243-266.

Brandvain Y, Kenney AM, Flagel L, Coop G, Sweigart AL. 2014. Speciation and introgression between *Mimulus nasutus* and *Mimulus guttatus*. PLoS Genet 10:e1004410.

Brodersen P, Petersen M, Bjørn Nielsen H, Zhu S, Newman M-A, Shokat KM, Rietz S, Parker J, Mundy J. 2006. *Arabidopsis* MAP kinase 4 regulates salicylic acid- and jasmonic acid/ethylene-dependent responses via EDS1 and PAD4. The Plant Journal 47:532-546.

Brown JW, Calixto CP, Zhang R. 2017. High-quality reference transcript datasets hold the key to transcript-specific RNA-sequencing analysis in plants. New Phytol 213:525-530.

Burrell AM, Hawkins AK, Pepper AE. 2012. Genetic analyses of nickel tolerance in a North American serpentine endemic plant, *Caulanthus amplexicaulis* var. *barbarae* (Brassicaceae). Am J Bot 99:1875-1883.

Burrell AM, No E-G, Pepper AE. 2011. Discovery of nuclear and plastid microsatellites, and other key genomic information, in the rare endemic plant (*Caulanthus amplexicaulis* var. *barbarae*) using minimal 454 pyrosequencing. Conservation Genetics Resources 3:753-755.

Burrell AM, Taylor KG, Williams RJ, Cantrell RT, Menz MA, Pepper AE. 2011. A comparative genomic map for *Caulanthus amplexicaulis* and related species (Brassicaceae). Molecular Ecology 20:784-798.

Bustos R, Castrillo G, Linhares F, Puga MI, Rubio V, Pérez-Pérez J, Solano R, Leyva A, Paz-Ares J. 2010. A Central Regulatory System Largely Controls Transcriptional Activation and Repression Responses to Phosphate Starvation in *Arabidopsis*. PLOS Genetics 6:e1001102.

Casal JJ. 2013. Photoreceptor signaling networks in plant responses to shade. Annu Rev Plant Biol 64:403-427.

Charlesworth B. 2009. Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. Nat Rev Genet 10:195-205.

Chaves I, Pokorny R, Byrdin M, Hoang N, Ritz T, Brettel K, Essen LO, van der Horst GT, Batschauer A, Ahmad M. 2011. The cryptochromes: blue light photoreceptors in plants and animals. Annu Rev Plant Biol 62:335-364.

Chen F, Mackey AJ, Vermunt JK, Roos DS. 2007. Assessing performance of orthology detection strategies applied to eukaryotic genomes. PLoS One 2.

Cheng MC, Kuo WC, Wang YM, Chen HY, Lin TP. 2017. UBC18 mediates ERF1 degradation under light-dark cycles. New Phytol 213:1156-1167.

Christensen A, Svensson K, Thelin L, Zhang W, Tintor N, Prins D, Funke N, Michalak M, Schulze-Lefert P, Saijo Y, et al. 2010. Higher Plant Calreticulins Have Acquired Specialized Functions in *Arabidopsis*. PLoS One 5:e11342.

Cifuentes-Esquivel N, Bou-Torrent J, Galstyan A, Gallemí M, Sessa G, Salla Martret M, Roig-Villanova I, Ruberti I, Martínez-García JF. 2013. The bHLH proteins BEE and BIM positively modulate the shade avoidance syndrome in *Arabidopsis* seedlings. The Plant Journal 75:989-1002.

Cokus SJ, Gugger PF, Sork VL. 2015. Evolutionary insights from de novo transcriptome assembly and SNP discovery in California white oaks. BMC Genomics 16:552.

Colosimo PF, Hosemann KE, Balabhadra S, Villarreal G, Dickson M, Grimwood J, Schmutz J, Myers RM, Schluter D, Kingsley DM. 2005. Widespread Parallel Evolution in Sticklebacks by Repeated Fixation of Ectodysplasin Alleles. Science 307:1928.

Conant GC, Wolfe KH. 2008. Turning a hobby into a job: how duplicated genes find new functions. Nat Rev Genet 9:938-950.

Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21:3674-3676.

Craciun AR, Meyer C-L, Chen J, Roosens N, De Groodt R, Hilson P, Verbruggen N. 2012. Variation in HMA4 gene copy number and expression among *Noccaea caerulescens* populations presenting different levels of Cd tolerance and accumulation. Journal of Experimental Botany 63:4179-4189.

Das PK, Geul B, Choi S-B, Yoo S-D, Park Y-I. 2011. Photosynthesis-dependent anthocyanin pigmentation in *Arabidopsis*. Plant Signaling & Behavior 6:23-25.

Dassanayake M, Oh DH, Haas JS, Hernandez A, Hong H, Ali S, Yun DJ, Bressan RA, Zhu JK, Bohnert HJ, et al. 2011. The genome of the extremophile crucifer *Thellungiella parvula*. Nat Genet 43:913-918.

De Gernier H, De Pessemier J, Xu J, Cristescu SM, Van Der Straeten D, Verbruggen N, Hermans C. 2016. A Comparative Study of Ethylene Emanation upon Nitrogen Deficiency in Natural Accessions of *Arabidopsis thaliana*. Front Plant Sci 7:70.

Devisetty UK, Covington MF, Tat AV, Lekkala S, Maloof JN. 2014. Polymorphism identification and improved genome annotation of *Brassica rapa* through Deep RNA sequencing. G3 (Bethesda) 4:2065-2078.

Dolan RW. 1995. The Rare, Serpentine Endemic *Streptanthus morrisonii* (Brassicaceae) Species Complex, Revisited Using Isozyme Analysis. Systematic Botany 20:338-346.

Drouin G, Daoud H, Xia J. 2008. Relative rates of synonymous substitutions in the mitochondrial, chloroplast and nuclear genomes of seed plants. Mol Phylogenet Evol 49:827-831.

Drumm-Herrel H, Mohr H. 1985. Photosensitivity of seedlings differing in their potential to synthesize anthocyanin. Physiologia Plantarum 64:60-66.

Drummond AJ, Suchard MA, Xie D, Rambaut A. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol 29:1969-1973.

Ellison CE, Hall C, Kowbel D, Welch J, Brem RB, Glass NL, Taylor JW. 2011. Population genomics and local adaptation in wild isolates of a model microbial eukaryote. Proceedings of the National Academy of Sciences 108:2831-2836.

Elmer KR, Fan S, Gunter HM, Jones JC, Boekhoff S, Kuraku S, Meyer A. 2010. Rapid evolution and selection inferred from the transcriptomes of sympatric crater lake cichlid fishes. Mol Ecol 19 Suppl 1:197-211.

Essigmann B, Güler S, Narang RA, Linke D, Benning C. 1998. Phosphate availability affects the thylakoid lipid composition and the expression of SQD1, a gene required for sulfolipid biosynthesis in *Arabidopsis thaliana*. Proceedings of the National Academy of Sciences 95:1950.

Eyre-Walker A, Keightley PD, Smith NGC, Gaffney D. 2002. Quantifying the Slightly Deleterious Mutation Model of Molecular Evolution. Molecular Biology and Evolution 19:2142-2149.

Feist LJ, Parker DR. 2001. Ecotypic variation in selenium accumulation among populations of *Stanleya pinnata*. New Phytologist 149:61-69.

Flagel LE, Wendel JF. 2009. Gene duplication and evolutionary novelty in plants. New Phytol 183:557-564.

Franzke A, Koch MA, Mummenhoff K. 2016. Turnip Time Travels: Age Estimates in Brassicaceae. Trends Plant Sci 21:554-561.

Freeman JL, Banuelos GS. 2011. Selection of salt and boron tolerant selenium hyperaccumulator *Stanleya pinnata* genotypes and characterization of Se phytoremediation from agricultural drainage sediments. Environ Sci Technol 45:9703-9710.

Goda H, Sawa S, Asami T, Fujioka S, Shimada Y, Yoshida S. 2004. Comprehensive comparison of auxin-regulated and brassinosteroid-regulated genes in *Arabidopsis*. Plant Physiol 134:1555-1573.

Gossmann TI, Song BH, Windsor AJ, Mitchell-Olds T, Dixon CJ, Kapralov MV, Filatov DA, Eyre-Walker A. 2010. Genome wide analyses reveal little evidence for adaptive evolution in many plant species. Mol Biol Evol 27:1822-1832.

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

Graur D, Zheng Y, Price N, Azevedo RB, Zufall RA, Elhaik E. 2013. On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE. Genome Biol Evol 5:578-590.

Guo H, Lee TH, Wang X, Paterson AH. 2013. Function relaxation followed by diversifying selection after whole-genome duplication in flowering plants. Plant Physiol 162:769-778.

Gururani MA, Venkatesh J, Tran LS. 2015. Regulation of Photosynthesis during Abiotic Stress-Induced Photoinhibition. Mol Plant 8:1304-1320.

Hamblin MT, Thompson EE, Di Rienzo A. 2002. Complex Signatures of Natural Selection at the Duffy Blood Group Locus. The American Journal of Human Genetics 70:369-383.

Hawkins AK, Garza ER, Dietz VA, Hernandez OJ, Hawkins WD, Burrell AM, Pepper AE. 2017. Transcriptome Signatures of Selection, Drift, Introgression, and Gene Duplication in the Evolution of an Extremophile Endemic Plant. Genome Biol Evol 9:3478-3494.

Hertweck KL, Kinney MS, Stuart SA, Maurin O, Mathews S, Chase MW, Gandolfo MA, Pires JC. 2015. Phylogenetics, divergence times and diversification from three genomic partitions in monocots. Botanical Journal of the Linnean Society 178:375-393.

Hohmann N, Wolf EM, Lysak MA, Koch MA. 2015. A Time-Calibrated Road Map of Brassicaceae Species Radiation and Evolutionary History. Plant Cell 27:2770-2784.

Howell J. 1962. New western plants IV. Leaflets of Western Botany 9:223-224.

Hughes AL. 2005. Gene duplication and the origin of novel proteins. Proc Natl Acad Sci U S A 102:8791-8792.

Initiative TAG. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. nature 408:796-815.

Ivalu Cacho N, Millie Burrell A, Pepper AE, Strauss SY. 2014. Novel nuclear markers inform the systematics and the evolution of serpentine use in Streptanthus and allies (Thelypodieae, Brassicaceae). Mol Phylogenet Evol 72:71-81.

Jantzen SG, Sutherland BJ, Minkley DR, Koop BF. 2011. GO Trimming: Systematically reducing redundancy in large Gene Ontology datasets. BMC Res Notes 4:267.

Jensen TH, Jacquier A, Libri D. 2013. Dealing with pervasive transcription. Mol Cell 52:473-484.

Jiang C, Gao X, Liao L, Harberd NP, Fu X. 2007. Phosphate starvation root architecture and anthocyanin accumulation responses are modulated by the gibberellin-DELLA signaling pathway in *Arabidopsis*. Plant Physiol 145:1460-1470.

Jones DA, Jones JDG. 1997. The Role of Leucine-Rich Repeat Proteins in Plant Defences. In: Andrews JH, Tommerup IC, Callow JA, editors. Advances in Botanical Research: Academic Press. p. 89-167.

Jones MB, Williams WA, Ruckman JE. 1977. Fertilization of *Trifolium subterraneum* L. Growing on Serpentine Soils1. Soil Science Society of America Journal 41:87-89.

Kagale S, Robinson SJ, Nixon J, Xiao R, Huebert T, Condie J, Kessler D, Clarke WE, Edger PP, Links MG, et al. 2014. Polyploid evolution of the Brassicaceae during the Cenozoic era. Plant Cell 26:2777-2791.

Kazakou E, Dimitrakopoulos PG, Baker AJ, Reeves RD, Troumbis AY. 2008. Hypotheses, mechanisms and trade-offs of tolerance and adaptation to serpentine soils: from species to ecosystem level. Biol Rev Camb Philos Soc 83:495-508.

Khozin-Goldberg I, Cohen Z. 2006. The effect of phosphate starvation on the lipid and fatty acid composition of the fresh water eustigmatophyte *Monodus subterraneus*. Phytochemistry 67:696-701.

Kimura M. 1984. The neutral theory of molecular evolution: Cambridge University Press.

Kondrashov FA. 2012. Gene duplication as a mechanism of genomic adaptation to a changing environment. Proc Biol Sci 279:5048-5057.

Koyama T, Nii H, Mitsuda N, Ohta M, Kitajima S, Ohme-Takagi M, Sato F. 2013. A regulatory cascade involving class II ETHYLENE RESPONSE FACTOR transcriptional repressors operates in the progression of leaf senescence. Plant Physiol 162:991-1005.

Kruckeberg AR. 1951. Intraspecific Variability in the Response of Certain Native Plant Species to Serpentine Soil. American Journal of Botany 38:408-419.

Kruckeberg AR, Rabinowitz D. 1985. Biological Aspects of Endemism in Higher Plants. Annual Review of Ecology and Systematics 16:447-479.

Kwok SF, Piekos B, Misera S, Deng XW. 1996. A Complement of Ten Essential and Pleiotropic Arabidopsis COP/DET/FUS Genes Is Necessary for Repression of Photomorphogenesis in Darkness. Plant Physiology 110:731-742.

Lahti DC, Johnson NA, Ajie BC, Otto SP, Hendry AP, Blumstein DT, Coss RG, Donohue K, Foster SA. 2009. Relaxed selection in the wild. Trends Ecol Evol 24:487-496.

Lee S-j, Cho D-i, Kang J-y, Kim SY. 2015. An ARIA-interacting AP2 domain protein is a novel component of ABA signaling. Molecules and Cells 27:409-416.

Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22:1658-1659.

Loehlin DW, Carroll SB. 2016. Expression of tandem gene duplicates is often greater than twofold. Proceedings of the National Academy of Sciences 113:5988-5992.

Lynch M, Conery JS. 2000. The Evolutionary Fate and Consequences of Duplicate Genes. Science 290:1151.

Makino T, Kawata M. 2012. Habitat variability correlates with duplicate content of *Drosophila* genomes. Mol Biol Evol 29:3169-3179.

Mayer MS, Beseda L. 2010. Reconciling Taxonomy And Phylogeny in the *Streptanthus glandulosus* Complex (Brassicaceae)1. Annals of the Missouri Botanical Garden 97:106-116.

Montoya-Burgos JI. 2011. Patterns of positive selection and neutral evolution in the protein-coding genes of *Tetraodon* and *Takifugu*. PLoS One 6:e24800.

Mugal CF, Wolf JB, Kaj I. 2014. Why time matters: codon evolution and the temporal dynamics of dN/dS. Mol Biol Evol 31:212-231.

Mummidi S, Ahuja SS, Gonzalez E, Anderson SA, Santiago EN, Stephan KT, Craig FE, O'Connell P, Tryon V, Clark RA, et al. 1998. Genealogy of the CCR5 locus and chemokine system gene variants associated with altered rates of HIV-1 disease progression. Nature medicine 4:786-793.

Neme R, Tautz D. 2016. Fast turnover of genome transcription across evolutionary time exposes entire non-coding DNA to de novo gene emergence. Elife 5:e09977.

Nielsen R. 2005. Molecular Signatures of Natural Selection. Annual Review of Genetics 39:197-218.

Oh DH, Hong H, Lee SY, Yun DJ, Bohnert HJ, Dassanayake M. 2014. Genome structures and transcriptomes signify niche adaptation for the multiple-ion-tolerant extremophyte *Schrenkiella parvula*. Plant Physiol 164:2123-2138.

Ohno S. 1970. Evolution by Gene Duplication: New York: Springer.

Ohta T. 1972. Population size and rate of evolution. Journal of Molecular Evolution 1:305-314.

Oono Y, Seki M, Satou M, Iida K, Akiyama K, Sakurai T, Fujita M, Yamaguchi-Shinozaki K, Shinozaki K. 2006. Monitoring expression profiles of *Arabidopsis* genes during cold acclimation and deacclimation using DNA microarrays. Functional & Integrative Genomics 6:212-234.

Parra G, Bradnam K, Korf I. 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics 23:1061-1067.

Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. 2017. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods 14:417-419.

Pepper AE, Norwood LE. 2001. Evolution of *Caulanthus amplexicaulis* var. *barbarae* (Brassicaceae), a rare serpentine endemic plant: a molecular phylogenetic perspective. American Journal of Botany 88:1479-1489.

Qiu Q, Zhang G, Ma T, Qian W, Wang J, Ye Z, Cao C, Hu Q, Kim J, Larkin DM, et al. 2012. The yak genome and adaptation to life at high altitude. Nat Genet 44:946-949.

Ramírez V, Coego A, López A, Agorio A, Flors V, Vera P. 2009. Drought tolerance in *Arabidopsis* is controlled by the OCP3 disease resistance regulator. The Plant Journal 58:578-591.

Raven PH, Axelrod DI. 1995. Origin and relationships of the California flora: California Native Plant Society.

Reeves RD, Macfarlane RM, Brooks RR. 1983. Accumulation of Nickel and Zinc by Western North American Genera Containing Serpentine-Tolerant Species. American Journal of Botany 70:1297-1303.

Regvar M, Vogel K, Irgel N, Wraber T, Hildebrandt U, Wilde P, Bothe H. 2003. Colonization of pennycresses (Thlaspi spp.) of the Brassicaceae by arbuscular mycorrhizal fungi. Journal of Plant Physiology 160:615-626.

Ren L, Tan XJ, Xiong YF, Xu K, Zhou Y, Zhong H, Liu Y, Hong YH, Liu SJ. 2014. Transcriptome analysis reveals positive selection on the divergent between topmouth culter and zebrafish. Gene 552:265-271.

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

Ross-Ibarra J, Wright SI, Foxe JP, Kawabe A, DeRose-Wilson L, Gos G, Charlesworth D, Gaut BS. 2008. Patterns of polymorphism and demographic history in natural populations of *Arabidopsis lyrata*. PLoS One 3:e2411.

Rubio V, Linhares F, Solano R, Martín AC, Iglesias J, Leyva A, Paz-Ares J. 2001. A conserved MYB transcription factor involved in phosphate starvation signaling both in vascular plants and in unicellular algae. Genes & Development 15:2122-2133.

Safford HD, Viers JH, Harrison SP. 2005. SERPENTINE ENDEMISM IN THE CALIFORNIA FLORA: A DATABASE OF SERPENTINE AFFINITY. Madroño 52:222-257.

Schlötterer C. 2015. Genes from scratch – the evolutionary fate of de novo genes. Trends in Genetics 31:215-219.

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

Shahollari B, Vadassery J, Varma A, Oelmüller R. 2007. A leucine-rich repeat protein is required for growth promotion and enhanced seed production mediated by the endophytic fungus Piriformospora indica in Arabidopsis thaliana. The Plant Journal 50:1-13.

Shapiro MD, Marks ME, Peichel CL, Blackman BK, Nereng KS, Jonsson B, Schluter D, Kingsley DM. 2004. Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. nature 428:717-723.

Sharma SB, Dixon RA. 2005. Metabolic engineering of proanthocyanidins by ectopic expression of transcription factors in *Arabidopsis thaliana*. Plant J 44:62-75.

Slotte T, Foxe JP, Hazzouri KM, Wright SI. 2010. Genome-wide evidence for efficient positive and purifying selection in *Capsella grandiflora*, a plant species with a large effective population size. Mol Biol Evol 27:1813-1821.

Smith BT, Klicka J. 2013. Examining the role of effective population size on mitochondrial and multilocus divergence time discordance in a songbird. PLoS One 8:e55161.

Srivastava A, Sarkar H, Gupta N, Patro R. 2016. RapMap: a rapid, sensitive and accurate tool for mapping RNA-seq reads to transcriptomes. Bioinformatics 32:i192-i200.

St Onge KR, Kallman T, Slotte T, Lascoux M, Palme AE. 2011. Contrasting demographic history and population structure in *Capsella rubella* and *Capsella grandiflora*, two closely related species with different mating systems. Mol Ecol 20:3306-3320.

Stebbins GL. 1942. THE GENETIC APPROACH TO PROBLEMS OF RARE AND ENDEMIC SPECIES. Madro 6:241-258.

Stracke R, Werber M, Weisshaar B. 2001. The R2R3-MYB gene family in *Arabidopsis thaliana*. Current Opinion in Plant Biology 4:447-456.

Strasburg JL, Kane NC, Raduski AR, Bonin A, Michelmore R, Rieseberg LH. 2011. Effective population size is positively correlated with levels of adaptive divergence among annual sunflowers. Mol Biol Evol 28:1569-1580.

Tamate SC, Kawata M, Makino T. 2014. Contribution of nonohnologous duplicated genes to high habitat variability in mammals. Mol Biol Evol 31:1779-1786.

Tiffin P, Hahn MW. 2002. Coding sequence divergence between two closely related plant species: *Arabidopsis thaliana* and *Brassica rapa* ssp. *pekinensis*. J Mol Evol 54:746-753.

Turitzin SN. 1982. Nutrient limitations to plant growth in a California serpentine grassland. American Midland Naturalist, 107, 95-99. 107:95-99.

Turner TL, Bourne EC, Von Wettberg EJ, Hu TT, Nuzhdin SV. 2010. Population resequencing reveals local adaptation of *Arabidopsis lyrata* to serpentine soils. Nat Genet 42:260-263.

Twyford AD, Ennos RA. 2012. Next-generation hybridization and introgression. Heredity (Edinb) 108:179-189.

Van Bel M, Proost S, Van Neste C, Deforce D, Van de Peer Y, Vandepoele K. 2013. TRAPID: an efficient online tool for the functional and comparative analysis of de novoRNA-Seq transcriptomes. Genome Biology 14:1-10.

Wagner GP, Kin K, Lynch VJ. 2012. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. Theory in Biosciences 131:281-285.

Wang XW, Luan JB, Li JM, Su YL, Xia J, Liu SS. 2011. Transcriptome analysis and comparison reveal divergence between two invasive whitefly cryptic species. BMC Genomics 12:458.

Wang Z, Gerstein M, Snyder M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10:57-63.

Warwick SI, Sauder CA, Mayer MS, Al-Shehbaz IA. 2009. Phylogenetic relationships in the tribes Schizopetaleae and Thelypodieae (Brassicaceae) based on nuclear ribosomal ITS region and plastidndhF DNA sequences. Botany 87:961-985.

Whittaker RH. 1954. The Ecology of Serpentine Soils. Ecology 35:258-288.

Wolf JB, Kunstner A, Nam K, Jakobsson M, Ellegren H. 2009. Nonlinear dynamics of nonsynonymous (dN) and synonymous (dS) substitution rates affects inference of selection. Genome Biol Evol 1:308-319.

Wright S. 1931. EVOLUTION IN MENDELIAN POPULATIONS. Genetics 16:97-159.

Xie D-X, Feys BF, James S, Nieto-Rostro M, Turner JG. 1998. COI1: An *Arabidopsis* Gene Required for Jasmonate-Regulated Defense and Fertility. Science 280:1091.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Computer applications in the biosciences : CABIOS 13:555-556.

Yang Z, Nielsen R. 2000. Estimating Synonymous and Nonsynonymous Substitution Rates Under Realistic Evolutionary Models. Molecular Biology and Evolution 17:32-43.

Yue JX, Yu JK, Putnam NH, Holland LZ. 2014. The transcriptome of an amphioxus, *Asymmetron lucayanum*, from the Bahamas: a window into chordate evolution. Genome Biol Evol 6:2681-2696.

Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18:821-829.

Zhang J. 2003. Evolution by gene duplication: an update. Trends in Ecology & Evolution 18:292-298.

Zhang B. and Horvath S. (2005). A General Framework for Weighted Gene Co-Expression Network Analysis, Statistical Applications in Genetics and Molecular Biology: Vol. 4: No. 1, Article 17 PMID: 16646834

Zhulidov PA, Bogdanova EA, Shcheglov AS, Vagner LL, Khaspekov GL, Kozhemyako VB, Matz MV, Meleshkevitch E, Moroz LL, Lukyanov SA, et al. 2004. Simple cDNA normalization using kamchatka crab duplex-specific nuclease. Nucleic Acids Research 32:e37-e37.

Zimmermann M, Clarke O, Gulbis JM, Keizer DW, Jarvis RS, Cobbett CS, Hinds MG, Xiao Z, Wedd AG. 2009. Metal binding affinities of *Arabidopsis* zinc and copper transporters: selectivities match the relative, but not the absolute, affinities of their amino-terminal domains. Biochemistry 48:11640-11654.

APPENDIX A

SUPPLEMENTAL TABLES FROM CHAPTER II

Supplemental Table 1. Plant growth conditions for RNA samples. Floral and silique samples from CAB where unavailable due to their delayed flowering time.

Species	Tissue type	Condition
Both	Young leaf	1/4x MS media
Both	Stem	1/4x MS media
CAA only	Floral	1/4x MS media
Both	Senescent leaf	1/4x MS media
Both	Root	1/4x MS media
Both	Caudex	1/4x MS media
CAA only	Silique	1/4x MS media
Both	Whole	1/4x MS media + 60 µM Ni
Both	Whole	1/4x MS media (155 µM KNO_3)
Both	Whole	1/4x MS media + 100 mM NaCl

Supplemental Table 2. Descriptive statistics on all representative transcript loci (RTLs). "No hits" refer to transcripts with no BLAST data to any published reference (e.g., nt/nr/TAIR).

	CAA	CAB
All (RTL)	93,647	83,484
Length of transcriptome (Mbp)	64.5	45.2
N50 or other metrics	1,001	691
With Blastn hit to nt	63,867	56,460
With Blastx hit to nr protein	69,132	66,642
With Blastx hit to core euk genes	248/248	248/248
No hit to any database (nt/nr)	19,696	17,613
Plastid transcripts	42	55
Mitochondrial transcripts	44	48
Total hits to TAIR nuclear CDS ORFs >99 bp	19,555	17,912

Supplemental Table 3. Descriptive statistics for TOPs from reciprocal blast hits (RBH).

Total TOPs	32,249
Total TOPs with ORFs (>99 bp)	30,643
TOPs with no hits to nt/nr	2,558
Coding sequence SNPs	53,359
Coding SNPs frequency	0.45%
Indels in ORFs	896
TOPs with frameshift	242
dN (mean/median)	0.0155/ 0.0070
dS (mean/median)	0.0615 / 0.0291
dN/dS (mean/median)	0.3458 / 0.2204