Transcriptome Signatures of Selection, Drift, Introgression, and Gene Duplication in the Evolution of an Extremophile Endemic Plant

Angela K. Hawkins[1], Elyssa R. Garza[1], Valerie A. Dietz[1], Oscar J. Hernandez[1], W. Daryl Hawkins[2], A. Millie Burrell[1], and Alan E. Pepper[1,*]

[1]Department of Biology, Texas A&M University

[2]Department of Nuclear Engineering, Texas A&M University

*Corresponding author: E-mail: apepper@bio.tamu.edu.

Accepted: December 4, 2017

Abstract

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, and Ca), as well as toxic levels of heavy metals (e.g., Ni, Cr, and 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 with 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 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, Thelypodieae, ultramafic, *Streptanthus*.

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, and 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 1954; 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 nonserpentine 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 very high rates of edaphic endemism (Whittaker 1954; 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, 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 ~75 km. 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 for a comparison of protein coding genes and to 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 nonselective 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 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, whereas the CAB1 line was obtained from a serpentine barren in Santa Barbara County, CA (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 Material online). As base media for manipulation of nutrient conditions, we used Murashige and Skoog (MS) medium, 1/4 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_079505 C 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_079505 C were grown in 1/4× MS media, as described, but with varying concentrations of KPO_4 for

28 days (*Caulanthus*) or 22 days (*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 (2×100 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 Illumina 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). 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., nonoverlapping) 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/; last accessed December 2017). 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 ten (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 with 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 and GGBZ00000000 CAA.

Functional Annotation

RTL were queried against the NCBI nonredundant 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 <5 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 nonsynonymous 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 Wolfe 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 (supplementary table S1, Supplementary Material online). 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 Illumina paired-end reads for *CAA1* and ~50 million paired-end reads for CAB1. Trimmed reads were assembled de novo and merged to yield a nonredundant set of 93,647 representative transcribed loci (RTL) for CAA1 (N50 = 1,001 bp), and 83,484 RTL for CAB1 (N50 = 691 bp) (supplementary table S2, Supplementary Material online). Approximately 68% of the RTL from both transcriptomes had significant BLASTN hits (*e*-value $<10^{-8}$) to sequences in the NCBI nonredundant 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 nonredundant 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) (supplementary table S3, Supplementary Material online). 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

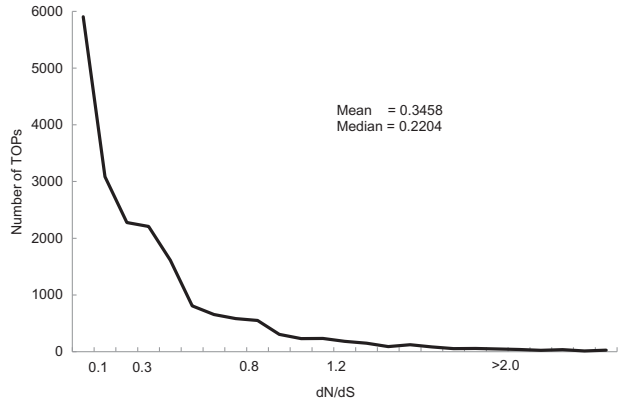

Fig. 1.—Histogram of d*N*/d*S* ratios of tentative orthologous pairs (TOPs) from CAA1 and CAB1 as calculated using the yn00 method in PAML.

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 d*N*/d*S* ratio of 0.346 (median = 0.220) (fig. 1). This value was relatively high compared with 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 d*N*/d*S* 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 S4*A*, Supplementary Material online). 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. 2).

In addition, there were 3,144 TOPs that had a $dN > 0$, but the d*N*/d*S* ratio was undefined because $dS = 0$. In most studies examining d*N*/d*S*, such genes are ignored. Given the relatively 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 nonsynonymous 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 (supplementary table S4*B*, Supplementary Material online).

Effects of Population Size

Although the finding of protein encoding genes with high d*N*/d*S* ratios is widely attributed to positive selection

Table 1

A Representative Sampling of Global Mean d*N*/d*S* Ratios from EST, Transcriptome, and Genome Sequencing Data Sets

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

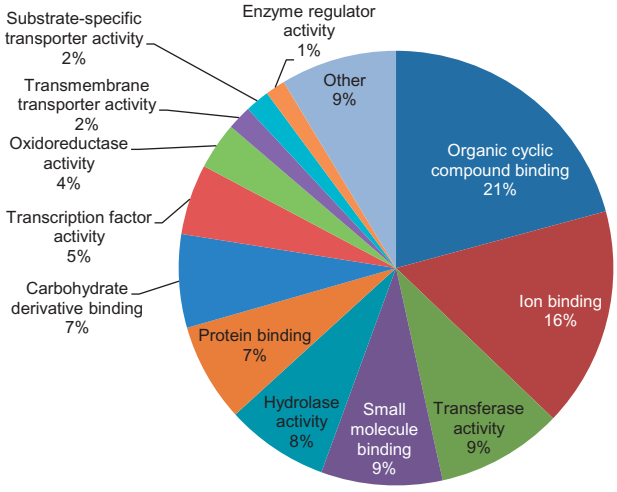

Fig. 2.—Distribution of level 3 gene ontology (GO) terms for molecular function among TOPs with d*N*/d*S* > 1.2.

(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 $<10^2$, 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 versus CAB1 comparison (0.238) was significantly higher ($P = 0.001$ in a Wilcoxon paired signed-rank test) than that of *A. lyrata* versus *C. grandiflora* (0.127). The median values differed similarly (0.136 for CAA1 vs. CAB1 and 0.069 for *A. lyrata* vs. *C. grandiflora*, $P = 0.0001$), Further, the elevated d*N*/d*S* values in CAA1 versus CAB1 occurred across a broad distribution of genes rather than in just the highest d*N*/d*S* categories (fig. 3). 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 versus CAB1 (rather than a few outliers with high d*N*/d*S*, as would be expected, e.g., 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.

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* $\sim$ 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). 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).

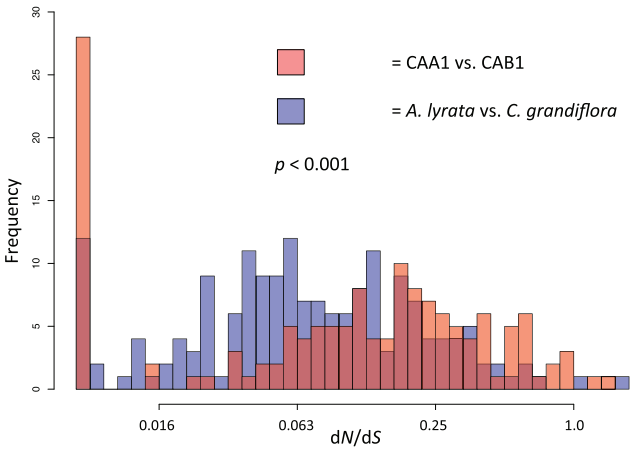

Fig. 3.—Overlay histogram comparing d*N*/d*S* ratios between CAA1 versus CAB1 (red) and *Capsella grandiflora* versus *Arabidopsis lyrata* (blue) based on 177 1:1:1:1 orthogroups. Column at far left corresponds to orthologs with a d*S* value of 0.

Table 2

d*N*/d*S* Values of CAA1 and CAB1 Genes in the Shade Avoidance Gene Ontology Category (GO: 0009641)

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 d*N*/d*S* ratio (ω_s).

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 overrepresented among loci with 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. 4). 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 with 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 nonoverlapping 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.

Transcription Factor 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 1954). 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*) 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 ~1.1 kb 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 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. 5*a*) 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 figure 5*b* and *c*, 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 2-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.

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) (supplementary table S4*C* and *D*, Supplementary Material online).

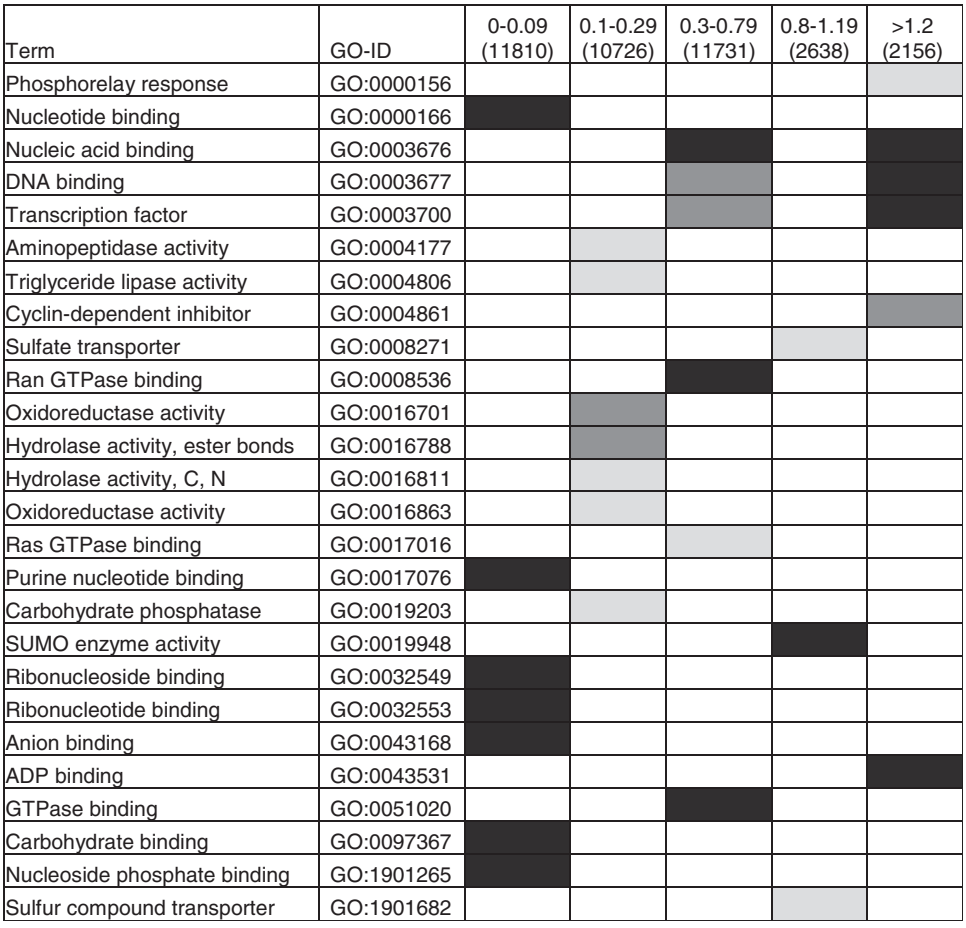

Fig. 4.—Heat map of enriched GO categories for TOPs based on dN/dS values from 0 to >1.2. Number of gene pairs in each category is indicated in parentheses. Colors indicate FDR values, light grey = 0.05, grey = 0.01, and black = 0.005.

Table 3

List of Enriched GO Terms in Transcripts with a Synthetic d*N*/d*S* > 1.2 as Compared with a Synthetic d*N*/d*S* < 1.2

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

Note.—Default FDR of 0.05 was implemented.

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 $dS = 8.4 \times 10^{-5}$, 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. 6 and 7). 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 data set 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

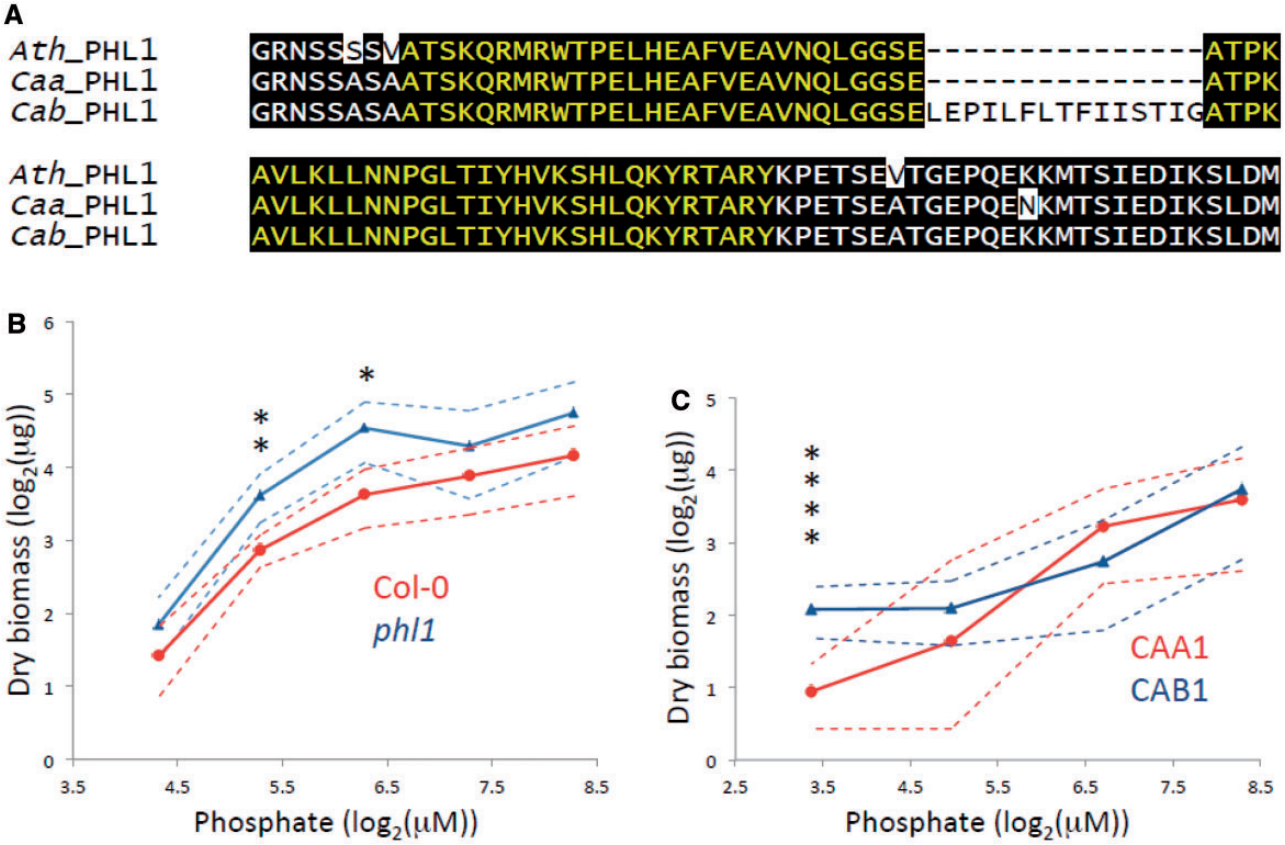

Fig. 5.—Phenotypes of loss-of-function variants in the MYB-CC transcription factor *PHL1*. (*a*) Protein alignment of *Ath*, *Caa*, and *Cab* PHL1. Yellow font indicates the conserved MYB DNA-binding domain. (*b*) Seedling growth phenotypes of *Ath* wild-type (Col-0) and *phl1* mutant in low phosphate conditions. Growth metric is mean end-point dry biomass measured 22 days 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 days after emergence of first true leaf.

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 the both 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. 6), 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% CI of 0.027–0.232 Ma), whereas analysis of alignments 24 randomly selected nuclear genes yielded an divergence time of 2.8 Ma (95% CI 2.3–3.2 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% CI of 2.8–3.6 Ma). From these results, we surmised that an initial divergence event occurred

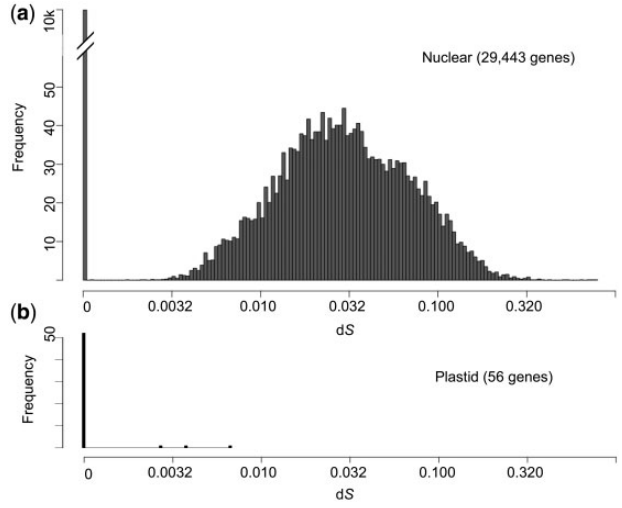

Fig. 6.—Rates of synonymous substitution (d*S*) between CAA1 and CAB1 orthologs. (*a*) Histogram of d*S* values for 29,443 nuclear genes. Column at the far left is comprised of orthologous pairs with d*S* = 0. (*b*) Histogram of d*S* values for 56 plastid encoded genes. Column at the far left corresponds to orthologous pairs with d*S* = 0. Nonzero 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.

~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 versus CAA1 and CAB1 versus CAB1 were used to identify 1,299 and 729 high-confidence tentative paralogous pairs (TPPs), respectively (supplementary table S3, Supplementary Material online). 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. 7) 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 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

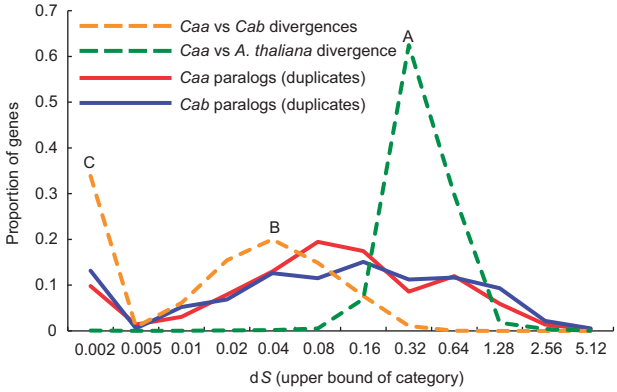

Fig. 7.—d*S* comparisons between orthologs (dashed lines) and paralogs (solid lines) as indicated. (*a*) The d*S* values of *Caa* versus *Ath* orthologs, with the major peak presumably corresponding to the Brassicaceae Lineage I/Lineage II (*Caa*/*Ath*) split ~23.4 Ma. (*b*) d*S* values of Caa versus Cab orthologs, with the right-most peak corresponding to the putative earlier divergence between *Caa* and *Cab* lineages (~3.2 Ma). (*c*) The much more recent divergence of *Caa* from *Cab* orthologs (~0.125 Ma).

Lineage I and Lineage II of the Brassicaceae family (fig. 7*a*) and is characterized by paralogs with a mean d*S* of ~0.8 (Guo et al. 2013; Kagale et al. 2014). Our data set 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. 7*b*), 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. 7*b*).

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. 8). 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. 8*b*), with very few genes exhibiting neutral evolution (i.e., d*N*/d*S*~1) in either taxon (fig. 8*a*). Thus, in this gene set, negative selection appears to have acted to retain two functional gene copies in both taxa.

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 with its serpentine intolerant sister taxon

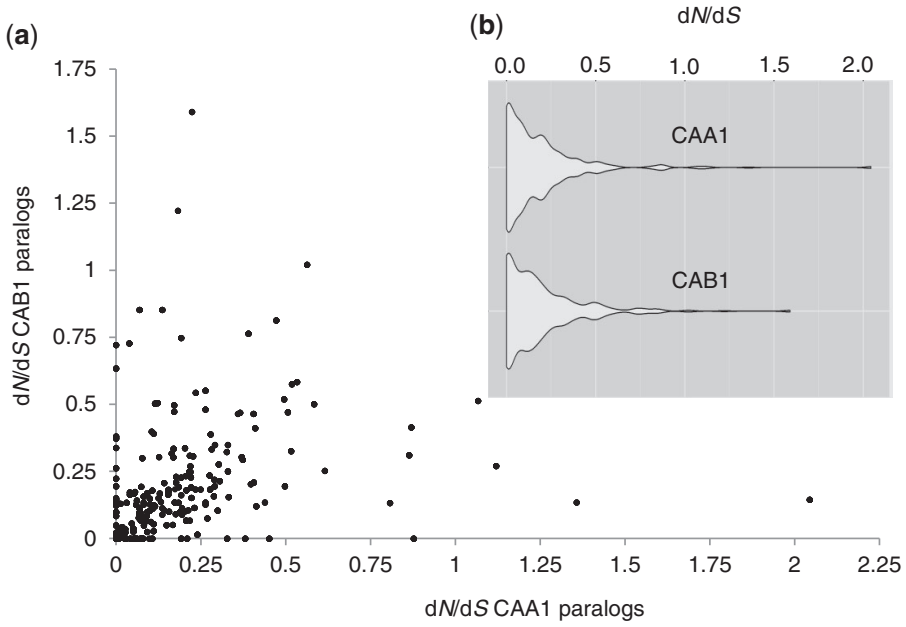

Fig. 8.—Coding sequence evolution in paralogous gene pairs that are shared between CAA1 and CAB1. (*a*) Scatterplot of d*N*/d*S* values of 194 corresponding parologous pairs. (*b*) Density distribution (violin) plots of d*N*/d*S* values for 194 parologous pairs within the CAA1 and CAB1 transcriptomes.

Caa. Although 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., $dN > 0$ and $dS \sim 0$) (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 $dS > 0.04$. 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* versus *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* versus *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. 2). 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, $\omega_s > 1.2$, and those loci with stop-codon polymorphisms (fig. 4, table 3, and supplementary table S4*B*, Supplementary Material online), 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, whereas 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 1954). 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 nonmonophyletic genera of *Caulanthus* and *Streptanthus* (Burrell, Taylor, et al. 2011; Cacho, Burrell, 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 (Burrell unpublished; 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 (Cacho, Burrell, et al. 2014), six nuclear markers and two plastid markers were used to estimate that serpentine tolerance arose independently ~4 times in the "ASHTB" subclade of the complex (credibility = 0.97) that includes *C. amplexicaulis*, the Sierra Nevada serpentine endemic *Streptanthus polygaloides*, *S. tortuosus* (which has serpentine tolerant and nontolerant ecotypes), a number of serpentine-tolerant *Streptanthus* species of the California Coastal Range, and several nonserpentine 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 (Cacho, Burrell, 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 (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. 7). 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; Cacho, Burrell, 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; Cacho, Burrell, 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; Cacho, Burrell, 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; Cacho, Burrell, 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 the 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.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

The authors gratefully acknowledge the use of computational resources provided by the Texas A&M Institute for Genome Sciences and Society (TIGSS). They thank S. Mandal, P. Greer, K. Kognathi, R. Perez, R. Aramayo, and T. McKnight for assistance with computational resources. Primary support for this project was from National Science Foundation (IOS) 12581020.

Literature Cited

Al-Shehbaz IA, Beilstein MA, Kellogg EA. 2006. Systematics and phylogeny of the Brassicaceae (Cruciferae): an overview. Plant Syst Evol. 259(2–4):89–120.

Alonso JM, et al. 2003. Genome-wide insertional mutagenesis of *Arabidopsis thaliana*. Science 301(5633):653.

Altenhoff AM, Dessimoz C. 2009. Phylogenetic and functional assessment of orthologs inference projects and methods. PLoS Comput Biol. 5(1):e1000262.
Anacker BL. 2014. The nature of serpentine endemism. Am J Bot. 101(2):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(4):e38.
Arnold BJ, et al. 2016. Borrowed alleles and convergence in serpentine adaptation. Pro Nat Aca Sci U S A. 113(29):8320–8325.
Bartish IV, et al. 2012. Phylogeny and colonization history of *Pringlea antiscorbutica* (Brassicaceae), an emblematic endemic from the South Indian Ocean Province. Mol Phylogenet Evol. 65(2):748–756.
Beilstein MA, Nagalingum NS, Clements MD, Manchester SR, Mathews S. 2010. Dated molecular phylogenies indicate a Miocene origin for *Arabidopsis thaliana*. Proc Natl Acad Sci U S A. 107(43): 18724–18728.
Blanc G, Wolfe KH. 2004. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. Plant Cell 16(7):1667–1678.
Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30(15):2114–2120.
Bou-Torrent J, Roig-Villanova I, Galstyan A, Martínez-García JF. 2008. PAR1 and PAR2 integrate shade and hormone transcriptional networks. Plant Signal Behav. 3(7):453–454.
Bowers JE, Chapman BA, Rong J, Paterson AH. 2003. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature 422(6930):433–438.
Brady KU, Kruckeberg AR, Bradshaw HD. 2005. Evolutionary ecology of plant adaptation to serpentine soils. Ann Rev Ecol Syst. 36(1):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(6):e1004410.
Brodersen P, et al. 2006. *Arabidopsis* MAP kinase 4 regulates salicylic acid- and jasmonic acid/ethylene-dependent responses via EDS1 and PAD4. Plant J. 47(4):532–546.
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(11):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. Conserv Genet Res. 3(4):753–755.
Burrell AM, Taylor KG, et al. 2011. A comparative genomic map for *Caulanthus amplexicaulis* and related species (Brassicaceae). Mol Ecol. 20(4):784–798.
Bustos R, et al. 2010. A central regulatory system largely controls transcriptional activation and repression responses to phosphate starvation in *Arabidopsis*. PLoS Genet. 6(9):e1001102.
Cacho IN, Burrell AM, 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.
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(3):195–205.
Chaves I, et al. 2011. The cryptochromes: blue light photoreceptors in plants and animals. Annu Rev Plant Biol. 62:335–364.
Christensen A, et al. 2010. Higher plant calreticulins have acquired specialized functions in *Arabidopsis*. PLoS One 5(6):e11342.
Cifuentes-Esquivel N, et al. 2013. The bHLH proteins BEE and BIM positively modulate the shade avoidance syndrome in *Arabidopsis* seedlings. Plant J. 75(6):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, et al. 2005. Widespread parallel evolution in sticklebacks by repeated fixation of ectodysplasin alleles. Science 307(5717):1928.
Conant GC, Wolfe KH. 2008. Turning a hobby into a job: how duplicated genes find new functions. Nat Rev Genet. 9(12):938–950.
Conesa A, et al. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21(18):3674–3676.
Craciun AR, et al. 2012. Variation in HMA4 gene copy number and expression among *Noccaea caerulescens* populations presenting different levels of Cd tolerance and accumulation. J Exp Bot. 63(11):4179–4189.
Dassanayake M, et al. 2011. The genome of the extremophile crucifer *Thellungiella parvula*. Nat Genet. 43(9):913–918.
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 4:2065–2078.
Dolan RW. 1995. The rare, serpentine endemic *Streptanthus morrisonii* (Brassicaceae) species complex, revisited using isozyme analysis. Syst Bot. 20(3):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(3):827–831.
Drummond AJ, Suchard MA, Xie D, Rambaut A. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol. 29(8):1969–1973.
Ellison CE, et al. 2011. Population genomics and local adaptation in wild isolates of a model microbial eukaryote. Proc Nat Aca Sci U S A. 108(7):2831–2836.
Elmer KR, et al. 2010. Rapid evolution and selection inferred from the transcriptomes of sympatric Crater Lake cichlid fishes. Mol Ecol. 19:197–211.
Eyre-Walker A, Keightley PD, Smith NGC, Gaffney D. 2002. Quantifying the slightly deleterious mutation model of molecular evolution. Mol Bio Evol. 19(12):2142–2149.
Feist LJ, Parker DR. 2001. Ecotypic variation in selenium accumulation among populations of *Stanleya pinnata*. New Phytol. 149(1):61–69.
Flagel LE, Wendel JF. 2009. Gene duplication and evolutionary novelty in plants. New Phytol. 183(3):557–564.
Franzke A, Koch MA, Mummenhoff K. 2016. Turnip time travels: age estimates in Brassicaceae. Trends Plant Sci. 21(7):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(22):9703–9710.
Goda H, et al. 2004. Comprehensive comparison of auxin-regulated and brassinosteroid-regulated genes in *Arabidopsis*. Plant Physiol. 134(4):1555–1573.
Gossmann TI, et al. 2010. Genome wide analyses reveal little evidence for adaptive evolution in many plant species. Mol Biol Evol. 27(8):1822–1832.
Grabherr MG, et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 29(7):644–652.
Graur D, et al. 2013. On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE. Genome Biol Evol. 5(3):578–590.

Gruenheit N, et al. 2012. Cutoffs and k-mers: implications from a transcriptome study in allopolyploid plants. BMC Genomics 13:92.
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(2):769–778.
Hamblin MT, Thompson EE, Di Rienzo A. 2002. Complex signatures of natural selection at the duffy blood group locus. Am J Hum Genet. 70(2):369–383.
Hertweck KL, et al. 2015. Phylogenetics, divergence times and diversification from three genomic partitions in monocots. Bot J Linn Soc. 178(3):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(10):2770–2784.
Howell J. 1962. New western plants IV. Leaflets West Bot. 9:223–224.
Hughes AL. 2005. Gene duplication and the origin of novel proteins. Proc Natl Acad Sci U S A. 102(25):8791–8792.
Initiative TAG. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408(6814):796–815.
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(4):473–484.
Jones MB, Williams WA, Ruckman JE. 1977. Fertilization of *Trifolium subterraneum* L. growing on serpentine soils. Soil Sci Soc Am J. 41(1):87–89.
Kagale S, et al. 2014. Polyploid evolution of the Brassicaceae during the Cenozoic era. Plant Cell 26(7):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(4):495–508.
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(1749): 5048–5057.
Kruckeberg AR. 1951. Intraspecific variability in the response of certain native plant species to serpentine soil. Am J Bot. 38(6): 408–419.
Kruckeberg AR, Rabinowitz D. 1985. Biological aspects of endemism in higher plants. Annu Rev Ecol Syst. 16(1):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 Physiol. 110(3):731–742.
Lahti DC, et al. 2009. Relaxed selection in the wild. Trends Ecol Evol. 24(9):487–496.
Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22(13):1658–1659.
Loehlin DW, Carroll SB. 2016. Expression of tandem gene duplicates is often greater than twofold. Proc Natl Acad Sci U S A. 113(21):5988–5992.
Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. Science 290(5494):1151.
Mandáková T, Li Z, Barker MS, Lysak MA. 2017. Diverse genome organization following 13 independent mesopolyploid events in Brassicaceae contrasts with convergent patterns of gene retention. Plant J. 91(1):3–21.
Makino T, Kawata M. 2012. Habitat variability correlates with duplicate content of *Drosophila* genomes. Mol Biol Evol. 29(10): 3169–3179.
Mayer MS, Beseda L. 2010. Reconciling taxonomy and phylogeny in the *Streptanthus glandulosus* complex (Brassicaceae)1. Ann Mo Bot Gard. 97(1):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(9):e24800.
Mugal CF, Wolf JB, Kaj I. 2014. Why time matters: codon evolution and the temporal dynamics of dN/dS. Mol Biol Evol. 31(1):212–231.
Mummidi S, et al. 1998. Genealogy of the CCR5 locus and chemokine system gene variants associated with altered rates of HIV-1 disease progression. Nat Med. 4(7):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. Ann Rev Genet. 39:197–218.
Oh DH, et al. 2014. Genome structures and transcriptomes signify niche adaptation for the multiple-ion-tolerant extremophyte *Schrenkiella parvula*. Plant Physiol. 164(4):2123–2138.
Ohno S. 1970. Evolution by gene duplication. New York: Springer.
Ohta T. 1972. Population size and rate of evolution. J Mol Evol. 1(4):305–314.
Oono Y, et al. 2006. Monitoring expression profiles of *Arabidopsis* genes during cold acclimation and deacclimation using DNA microarrays. Funct Integr Genomics 6(3):212–234.
Parra G, Bradnam K, Korf I. 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics 23(9):1061–1067.
Pepper AE, Norwood LE. 2001. Evolution of *Caulanthus amplexicaulis* var. *barbarae* (Brassicaceae), a rare serpentine endemic plant: a molecular phylogenetic perspective. Am J Bot. 88(8):1479–1489.
Qiu Q, et al. 2012. The yak genome and adaptation to life at high altitude. Nat Genet. 44(8):946–949.
Ramírez V, et al. 2009. Drought tolerance in *Arabidopsis* is controlled by the OCP3 disease resistance regulator. Plant J. 58(4):578–591.
Raven PH, Axelrod DI. 1995. Origin and relationships of the California flora. Sacramento (CA): California Native Plant Society Press.
Ren L, et al. 2014. Transcriptome analysis reveals positive selection on the divergent between topmouth culter and zebrafish. Gene 552(2):265–271.
Ross-Ibarra J, et al. 2008. Patterns of polymorphism and demographic history in natural populations of *Arabidopsis lyrata*. PLoS One 3(6):e2411.
Rubio V, et al. 2001. A conserved MYB transcription factor involved in phosphate starvation signaling both in vascular plants and in unicellular algae. Genes Dev. 15(16):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 Genet. 31(4):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(8):1086–1092.
Shapiro MD, et al. 2004. Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. Nature 428(6984):717–723.
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(8):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(2):e55161.

St Onge KR, Källman T, Slotte T, Lascoux M, Palmé 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(16):3306–3320.
Stebbins GL. 1942. The genetic approach to problems of rare and endemic species. Madroño 6:241–258.
Strasburg JL, et al. 2011. Effective population size is positively correlated with levels of adaptive divergence among annual sunflowers. Mol Biol Evol. 28(5):1569–1580.
Tamate SC, Kawata M, Makino T. 2014. Contribution of nonohnologous duplicated genes to high habitat variability in mammals. Mol Biol Evol. 31(7):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(6):746–753.
Turitzin SN. 1982. Nutrient limitations to plant growth in a California serpentine grassland. Am Midl Nat. 107(1):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(3):260–263.
Twyford AD, Ennos RA. 2012. Next-generation hybridization and introgression. Heredity (Edinb) 108(3):179–189.
Van Bel M, et al. 2013. TRAPID: an efficient online tool for the functional and comparative analysis of de novoRNA-Seq transcriptomes. Genome Biol. 14(12):1–10.
Wang XW, et al. 2011. Transcriptome analysis and comparison reveal divergence between two invasive whitefly cryptic species. BMC Genomics 12:458.
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 plastid ndhF DNA sequences. Botany 87(10):961–985.
Whittaker RH. 1954. The ecology of serpentine soils. Ecology 35(2):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(0):308–319.
Wright S. 1931a. Evolution in Mendelian populations. Genetics 16(2):97–159.
Wright S. 1931b. Evolution in Mendelian populations. Genetics 16:63.
Xie D-X, Feys BF, James S, Nieto-Rostro M, Turner JG. 1998. An *Arabidopsis* gene required for jasmonate-regulated defense and fertility. Science 280(5366):1091.
Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci. 13(5):555–556.
Yang Z, Nielsen R. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol Biol Evol. 17(1):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(10):2681–2696.
Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18(5):821–829.
Zhang J. 2003. Evolution by gene duplication: an update. Trends Ecol Evol. 18(6):292–298.
Zhulidov PA, Bogdanova EA, Shcheglov AS, Vagner LL, Khaspekov GL, Kozhemyako VB, Matz MV, Meleshkevitch E, Moroz LL, Lukyanov SA. 2004. Simple cDNA normalization using Kamchatka crab duplex-specific nuclease. Nucleic Acids Res. 32:e37–e37.
Zimmermann M, et al. 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(49):11640–11654.

Associate editor: Bill Martin