RESEARCH ARTICLE

American Journal of Botany

INVITED SPECIAL ARTICLE

For the Special Issue: Using and Navigating the Plant Tree of Life

Molecular evolution of key metabolic genes during transitions to C_4 and CAM photosynthesis

Eric W. Goolsby[1,2,5], Abigail J. Moore[1,3], Lillian P. Hancock[1], Jurriaan M. De Vos[1,4], and Erika J. Edwards[1,2]

Manuscript received 19 September 2017; revision accepted 12 December 2017.

[1] Brown University, Department of Ecology and Evolutionary Biology, Box G-W, 80 Waterman St, Providence, Rhode Island 02912, USA

[2] Yale University, Department of Ecology and Evolutionary Biology, 165 Prospect Street, New Haven, Connecticut 06511, USA

[3] University of Oklahoma, Department of Microbiology and Plant Biology and Oklahoma Biological Survey, 136 George Lynn Cross Hall, 770 Van Vleet Oval, Norman, Oklahoma 73019, USA

[4] University of Basel, Department of Environmental Sciences–Botany, Bernoullistrasse 32, 4056, Basel, Switzerland

[5] Author for correspondence (e-mail: eric.goolsby.evolution@gmail.com)

Citation: Goolsby, E. W., A. J. Moore, L. P. Hancock, J. M. De Vos, and E. J. Edwards. 2018. Molecular evolution of key metabolic genes during transitions to C_4 and CAM photosynthesis. *American Journal of Botany* 105(3): 602–613.

doi:10.1002/ajb2.1051

PREMISE OF THE STUDY: Next-generation sequencing facilitates rapid production of well-sampled phylogenies built from very large genetic data sets, which can then be subsequently exploited to examine the molecular evolution of the genes themselves. We present an evolutionary analysis of 83 gene families (19 containing carbon-concentrating mechanism (CCM) genes, 64 containing non-CCM genes) in the portullugo clade (Caryophyllales), a diverse lineage of mostly arid-adapted plants that contains multiple evolutionary origins of all known photosynthesis types in land plants (C_3, C_4, CAM, C_4-CAM, and various intermediates).

METHODS: We inferred a phylogeny of 197 individuals from 167 taxa using coalescent-based approaches and individual gene family trees using maximum likelihood. Positive selection analyses were conducted on individual gene family trees with a mixed effects model of evolution (MEME). We devised new indices to compare levels of convergence and prevalence of particular residues between CCM and non-CCM genes and between species with different photosynthetic pathways.

KEY RESULTS: Contrary to expectations, there were no significant differences in the levels of positive selection detected in CCM versus non-CCM genes. However, we documented a significantly higher level of convergent amino acid substitutions in CCM genes, especially in C_4 taxa.

CONCLUSIONS: Our analyses reveal a new suite of amino acid residues putatively important for C_4 and CAM function. We discuss both the advantages and challenges of using targeted enrichment sequence data for exploratory studies of molecular evolution.

KEY WORDS Caryophyllales; convergent evolution; gene duplications; phylogeny; photosynthesis; portullugo; Portulacineae; positive selection.

The science of systematics has witnessed plenty of revolution within the last decades, in terms of both the data and the methods used for phylogenetic inference (e.g., Jukes and Cantor, 1969; Felsenstein, 1981, 1985; Hillis, 1987; Yang, 1996; Maddison, 1997; Lewis, 2001; Drummond et al. 2006; Degnan and Rosenberg, 2009). With the advent of next-generation sequencing (NGS), the landscape of phylogenetic systematics has shifted yet again, with both new challenges and opportunities presented by our growing ability to quickly amass genomic-scale sequence data across 100s or even 1000s of taxa (reviewed, e.g., by Straub et al., 2012; McCormack et al., 2013). The benefits of NGS to phylogenetic inference are obvious and were the first to be explored (Chaw et al., 2004; Dunn et al., 2008; Jiao et al., 2011; Wickett et al., 2014), but there is also enormous potential for the combination of large-scale genomic data with well-sampled and well-resolved phylogenies to inform many problems of molecular and phenotypic evolution. The most commonly used NGS approach to combine phylogenetic inference with studies of molecular evolution has been RNA sequencing (RNA-seq), or transcriptome analysis (e.g., Jiao et al., 2011; Wickett et al., 2014; Yang et al., 2015), and these efforts have mostly been focused on reconstructing patterns of gene duplication and loss through time. Another increasingly popular NGS method, targeted gene enrichment, has been successfully employed in phylogenetic inference (e.g., Faircloth et al., 2012; Lemmon et al., 2012; Mandel et al., 2015; Schmickl et al., 2016), but the accumulated data sets have hardly been explored in other contexts (but see Nevado et al., 2016). This lack of exploration may be due in part to the near exclusive use of putatively single copy loci (SCL) in bait design, which alleviates much of the difficulty in homology assignment of sequenced contigs. At the same time, however, focusing on SCL excludes many functionally important gene families from being sequenced, which in turn limits the use of these data sets for understanding the molecular evolution of most genes that might be involved in complex or interesting phenotypes of the focal clade.

Recently, we built a bioinformatics pipeline designed specifically to handle paralog sorting of large gene families sequenced with targeted gene enrichment (Moore et al., 2017), which then allowed us to design our gene sampling to include multiple, large gene families of functional interest. Our study lineage is the "portullugo" (Caryophyllales), a clade of ~2200 species of mostly arid-adapted succulent plants (Edwards and Ogburn, 2012). We are developing this group as a model lineage for studying the dynamics of photosynthesis evolution and ecological adaptation, as the clade harbors multiple origins of two alternative photosynthetic metabolisms in plants, C_4 and CAM photosynthesis (e.g., Edwards and Donoghue, 2006; Edwards and Ogburn, 2012; Ogburn and Edwards, 2013; Christin et al., 2011, 2014; Thulin et al., 2016; Moore et al., 2017). Briefly, C_4 and CAM are characterized by complex biochemical and anatomical alterations relative to the ancestral C_3 photosynthetic pathway that produce elevated levels of CO_2 inside photosynthetic organs, eliminating photorespiration and improving water-use efficiency (Fig. 1). In both C_4 and CAM, CO_2 is fixed by an enzyme found in all plants called phospho*enol*pyruvate carboxylase (PEPC) and converted into a 4-carbon acid. In C_4 plants, the acid is typically transported to the bundle sheath cells, where CO_2 is released and maintained at high concentrations so that ribulose-1,5-bisphosphate (RuBP) carboxylase/oxygenase (RuBisCO) preferentially reacts with CO_2, rather than O_2. Thus, C_4 plants utilize spatial separation of initial carbon capture and the Calvin cycle to avoid photorespiration. On the other hand, CAM plants employ temporal rather than spatial separation: at night, CO_2 is captured and accumulated as an acid; during the day, CO_2 is released for the Calvin cycle, thus allowing the stomata to remain closed during the day to prevent water loss. Thus, the two adaptations are united by a shared biochemical pathway, but are distinct in how they have isolated RuBisCO and created an internally elevated CO_2 environment. What makes these syndromes an even more elegant evolutionary study system is that, though fairly complex, both of these carbon-concentrating mechanisms have evolved hundreds of times throughout the last 30 Myr, making them two of the most convergent ecological adaptations in angiosperms (Sage et al., 2011; Edwards and Ogburn, 2012).

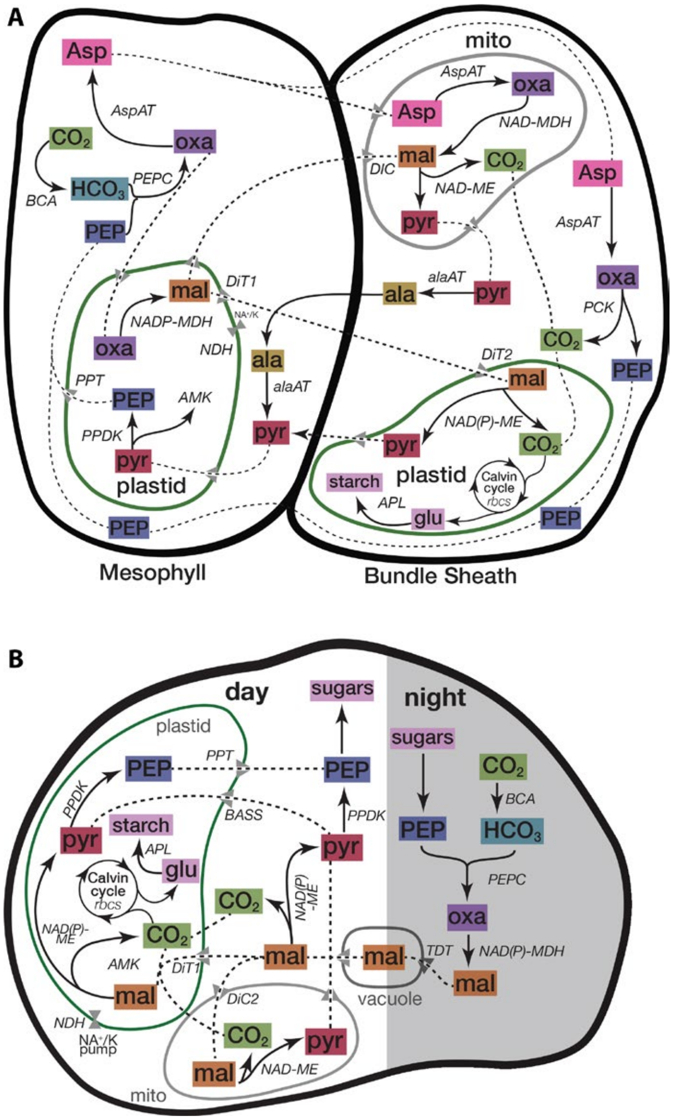

FIGURE 1. Schematic of carbon assimilation (A) via the C_4 photosynthetic pathway and (B) via the CAM photosynthetic pathway. The 19 gene families that code for major proteins of C_4/CAM biochemistry are italicized. Metabolites are represented with colored boxes. Protein abbreviations: alaAT = alanine aminotransferase, AMK = adenosine monophosphate-activated protein kinase, APL = glucose-1-phosphate adenylyltransferase, AspAT = aspartate aminotransferase, BASS = ile acid sodium symporter, BCA = beta-carbonic anhydrase, DIC = dicarboxylate carrier, DiT = dicarboxylate transporter, NAD-MDH = NAD malate dehydrogenase, NAD-ME = NAD malic enzyme, NADP-MDH = NADP malate dehydrogenase, NAD(P)-ME = NADP malic enzyme, NDH = NAD(P) H-plastoquinone-oxidoreducta, PCK = phosphoenolpyruvate carboxykinase, PEPC = phosphoenolpyruvate carboxylase, PPDK = phosphate dikinase, PPT = phosphoenolpyruvate-phosphate translocator, RbcS = ribulose bisphosphate carboxylase small chain, TDT = tonoplast dicarboxylate transporter.

Understanding the evolution of a full C_4 or CAM metabolism requires piecing together the ecological, anatomical, biochemical, and genetic aspects of theses syndromes and reconstructing the evolutionary order in their assembly. Here, we focus exclusively on one piece of the puzzle. Our goals were to design a set of enrichment probes (Moore et al., 2017) that would work well for phylogenetic analysis across the portullugo, a large lineage spanning ~50 Myr of evolution, and that would also permit us to amass a large sequence database of 19 key gene families that code for major proteins of

C_4/CAM biochemistry (hereafter referred to as carbon-capturing mechanism (CCM) genes; Fig. 1). All of these proteins are produced in all plants and belong to large gene families with complicated histories of duplication and loss. It is generally poorly understood how particular paralogs have been recruited into their new photosynthetic function and how these new functions may provide strong selection for adaptive evolution at the DNA sequence level.

To date, gene recruitment and adaptive protein evolution have been best studied in C_4 origins, rather than CAM, and in particular in C_4 grasses, which include roughly one third of the known origins of C_4 photosynthesis (Grass Phylogeny Working Group II, 2012; Sage et al., 2011). Both grasses and members of Caryophyllales demonstrate significant bias in what gene copy is recruited into C_4 function (Christin et al., 2013, 2015) and are also remarkable for convergent and adaptive evolution at multiple residues in the major C_4/CAM enzyme, PEPC (Christin et al., 2007). In the emerging model of C_4 evolution, the appearance of a C_4-like leaf anatomy is typically the first step (Christin et al., 2013), which is then followed by regulatory changes in gene expression (Sage et al., 2012; Williams et al., 2013). While adaptive evolution of the CCM proteins themselves does not always appear to have occurred (Lara et al., 2006; Silvera et al., 2014), positive selection and convergent evolution has been detected in PEPC sequences in species representing multiple C_4/CAM origins and is considered a final "optimization" step of fine-tuning a fully formed C_4/CAM system (Christin et al., 2012, 2014; Dunning et al., 2017).

Here we examine the molecular evolution of C_4 and CAM photosynthesis within the portullugo clade of Caryophyllales, extending our gene sampling from PEPC to those encoding 18 additional key enzymes and transporters of C_4 and CAM biochemical cycles (Fig. 1). We approach this system by testing for positive selection at specific codon positions and by assessing both the levels of molecular convergence within and between C_4 and CAM lineages, as well as the relative prevalence of amino acids in specific residues relative to photosynthetic pathways. In addition to containing C_3, CAM, and C_4 species, the portullugo clade contains many predominantly C_3 species with varying levels of low-level or inducible CAM activity (Guralnick and Jackson, 2001; Winter and Holtum, 2014; Holtum et al., 2017a, b), hereafter collectively referred to as C_3-CAM species. By considering the full continuum of C_3, CAM, and C_4 evolutionary trajectories, we aim to identify where along the trajectory most molecular adaptation may be occurring. Together, these exploratory approaches provide us with several novel directions for further study. By assessing amino acid substitutions that have evolved independently multiple times in C_4, CAM, as well as C_3-CAM lineages, and by assessing specific amino acid distributions prevalent within C_4 or CAM lineages that are sparse or absent in C_3 lineages, we identify multiple new residues putatively relevant to enzymatic function in C_4 and CAM metabolism.

MATERIALS AND METHODS

Taxon sampling

Expanding on the sampling from Moore et al. (2017) and Hancock et al. (in revision), which together included 142 taxa, we sequenced 55 additional individuals. Our final sampling includes 197 individuals representing 167 taxa and the major lineages of portullugo. Additionally, transcriptomes from *Pereskia bleo* DC. and *Portulaca oleracea* L. were included, as well as five non-portullugo transcriptomes (*Amaranthus hypochondriacus* L., *Boerhavia coccinea* Mill., *Mesembryanthemum crystallinum* L., *Trianthema portulacastrum* L., and *Beta vulgaris* L.) and six non-Caryophyllales model plants (*Arabidopsis thaliana* Schur, *Vitis vinifera* L., *Populus trichocarpa* Torr. & A.Gray, *Glycine max* Merr., *Oryza sativa* L., and *Solanum lycopersicum* L.), for a total of 197 individuals and 167 taxa (1KP; Matasci et al., 2014), as listed in Appendix S1 (see Supplemental Data with this article). Non-Caryophyllales were used to distinguish long branches separating deep gene duplications within Caryophyllales, but were not considered for downstream phylogenetic inference or for selection analyses.

Probe design, sequencing, and software pipeline

For a full description of methods, we refer to Moore et al. (2017). Briefly, we designed targeted enrichment probes (originally designed for Moore et al., 2017) based on portullugo transcriptomes from our previous work (Christin et al., 2014, 2015; *Anacampseros filamentosa* (Haw.) Sims, *Echinocereus pectinatus* Engelm., *Nopalea cochenillifera* (L.) Salm-Dyck, *Pereskia bleo*, *Pereskia grandifolia* Haw., *Pereskia lychnidiflora* DC., *Portulaca oleracea*, and *Talinum portulacifolium* Asch. ex Schweinf.) and from four species within Molluginaceae from the 1000 Plants transcriptome sequencing project [1KP; Matasci et al., 2014; *Hypertelis cerviana* (L.) Thulin (named *M. cerviana* Ser. in 1KP], *Mollugo verticillata* L., *Paramollugo nudicaulis* (Lam.) Thulin (named *M. nudicaulis* Lam. in 1KP), and *Trigastrotheca pentaphylla* (L.) Thulin (named *M. pentaphylla* L. in 1KP)). We designed MyBaits probes from 19 a priori designated CAM and C_4 associated (i.e., CCM) gene families, as well as 53 additional nuclear genes with *Arabidopsis* homologues and low to moderate copy numbers (MYcroarray, Ann Arbor, MI, USA). Gene families were matched to those identified by Christin et al. (2014, 2015), via BLAST (BLASTN 2.2.25, default settings; Altschul et al., 1990) against orthologous sequences of known identity from the non-Caryophyllales model species (Ensembl; Kersey et al., 2016).

Genomic DNA was extracted using the FastDNA Spin Kit (MP Biomedicals, Santa Ana, CA, USA), and samples were cleaned using QIAquick PCR CleanupKit (Qiagen, Valencia, CA, USA). Extracted DNA was sonicated and libraries were prepared using the NEBNext Ultra or NEBNext Ultra II DNA Library Prep Kits for Illumina (New England Biolabs, Ipswich, MA, USA). Sequencing was performed at the Brown University Genomics Core Facility or the Oklahoma Medical Research Foundation genomic sequencing facility on an Illumina HiSeq 2000 or 2500, and reads were submitted to the NCBI SRA (BioProjects PRJNA387599, PRJNA417446, and PRJNA415977; accession numbers in Appendix S1 with Supplemental Data).

Our pipeline reconstructs gene sequences by extracting short reads and assembling them into contigs, constructing longer sequences and assigning them to specific paralogs within gene families, and identifying within-family gene duplications and extracting phylogenetically informative orthologs (for detailed explanation, see Moore et al., 2017; scripts are available at https://github.com/abigail-Moore/baits-analysis). We subsequently reconstructed a multispecies coalescent phylogeny using ASTRAL II version 4.10.2 (Mirarab and Warner, 2015; Sayyari and Mirarab, 2016), after inferring gene trees using RAxML version 8 (Stamatakis, 2014).

Designation of photosynthetic pathway

For the purpose of identifying associations between molecular variation and photosynthetic pathway, we classified all included species into particular photosynthetic phenotypes. In some cases, the classifying was straightforward, particularly for well-known C_4 plants such as *Hypertelis cerviana* (Christin et al., 2011), or obvious constitutive CAM plants such as those found in the core cacti (Gibson and Nobel, 1986). However, the majority of portullugo species in our particular sample fell into neither of these categories; most species occupy phenotypic space somewhere between a "C_3 only" plant and a "full CAM" plant, in that they use C_3 metabolism as their primary method of carbon fixation, but also have a functional CAM biochemical cycle that either functions at low levels in conjunction with C_3, or can be upregulated in response to stress. We think these must be critical phenotypes along the C_3-CAM evolutionary trajectory, preceding the emergence of full CAM metabolism. The problem is that sorting these "low-level", C_3-CAM-like behaviors into rational categorical states is still an area of active debate (e.g., Winter et al., 2015), and furthermore, it is far more difficult to identify low-level CAM behavior than it is full CAM or C_4, so many species are still not definitively phenotyped. Thus, we classified our included species into C_3, low-level and/or facultative CAM (referred to hereafter as C_3-CAM), full CAM (referred to hereafter as CAM), and C_4. We based our designations on a combination of (1) published literature (Winter, 1979; Martin and Wallace, 2000; Guralnick and Jackson, 2001; Guralnick et al., 2008; Ocampo and Columbus, 2010; Winter and Holtum, 2014; Holtum et al., 2017a, b), (2) our own unpublished physiological data (i.e., drought experiments, isotope surveys), (3) plant morphology (succulence, or lack thereof) in observed specimens, (4) known habitat, and in several cases, (5) phylogenetic proximity to other taxa with a known photosynthetic pathway (Fig. 2). Species for which we had no concrete physiological data but could make reasonable assumptions based on categories 3–5 are highlighted in Fig. 2. Several species we have never observed in the field and lack physiological data, and these we did not attempt to phenotype but removed them from the convergence/prevalence analyses. Although three species, *Portulaca cryptopetala* Speg., *Hypertelis spergulacea* E.Mey. ex Fenzl, and *Mollugo verticillata*, are sometimes considered "C_3-C_4" species, they do not have an active C_4 biochemical pathway, and we therefore included them as C_3 or C_3-CAM (for *P. cryptopetala*) species here. Finally, we classified all other *Portulaca* as C_4 plants, though this unique lineage is also known to engage in a facultative CAM cycle (Koch and Kennedy, 1980; Guralnick and Jackson, 2001; Lara et al., 2004; Guralnick et al., 2008; Christin et al., 2014; Holtum et al., 2017a, b). A list of all photosynthetic pathway designations can be found in Appendix S2.

Identification of sites under positive selection

Gene family alignments and trees were analyzed using the HyPhy package (Kosakovsky Pond and Muse, 2005). We used a mixed effects model of evolution (MEME) to identify sites putatively under positive selection (for residue numbering, see Appendix S3) (Murrell et al., 2012). Similar to branch-site models (Yang and Reis, 2011), MEME identifies sites by statistically assessing whether ω, the ratio of nonsynonymous to synonymous substitutions, is significantly greater than one. Unlike the familiar branch-site model, which requires a priori designation of "background" and "foreground" branches, MEME allows ω to vary across sites as a fixed effect, while allowing for ω to vary from branch to branch at individual sites as a random effect that is marginalized in the likelihood calculation for each branch-site combination. Simulations have demonstrated that in the case of episodic diversifying positive selection, the a priori assignment of background and foreground branches is overly restrictive and may substantially increase the incidence of type I error, whereas the simultaneous detection of branches and sites under selection has been shown to be statistically unidentifiable (Kosakovsky Pond et al., 2011; Murrell et al., 2012).

As discussed above, the extent of C_4 and CAM pathway characteristics exhibited in many of our taxa are somewhat uncertain. Accordingly, the lack of a priori designations and associated parameter flexibility implemented in using the mixed effects model in MEME should improve our statistical power to identify amino acid residues under positive selection and simultaneously reduce the probability of type I error (Murrell et al., 2012). Because MEME performs independent hypothesis tests for each residue in a gene family, we applied false discovery rate corrections to site-specific p-values and set a nominal significance threshold of 0.05. Although less conservative than familywise error rate corrections, false discovery rate correction balances the risk of false positives while preserving statistical power (Benjamini and Hochberg, 1995; Murrell et al., 2012).

By running analyses on whole gene family trees, it is possible that a signal of positive selection could be "swamped out", either due to other phenotypes being too highly represented (e.g., there is selection in C_4 lineages but not in any others) or due to too many irrelevant paralogs included (e.g., there are five copies of *PPC*, but due to a strong gene recruitment bias only one paralog is ever performing a C_4/CAM function). To evaluate these issues, we also performed positive selection tests on subsets of our data where we pruned gene family trees to only include one photosynthetic phenotype: a C_3-only analysis, C_3-CAM-only, CAM-only, and C_4-only. For the gene families where we have identified the paralogs that have been recruited to C_4/CAM (Christin et al., 2015), we did a second analysis where we analyzed each paralog independently and compared the levels of selection between paralogs (Christin et al., 2015).

Identification of additional sites putatively relevant to C_4 and CAM photosynthesis

In addition to positive selection analyses, we identified molecular convergence and prevalence of specific amino acids that appear to be associated with C_4 or CAM lineages (Fig. 3). Specifically, amino acids that arose independently multiple times in non-C_3 lineages but did not originate from C_3 lineages may be of functional importance to C_4 or CAM photosynthetic pathways under convergent selective forces. We defined molecular convergence as more than one likely independent origin of an amino acid in C_3-CAM, CAM, or C_4 species with no independent origins in C_3 lineages. We estimated the number of independent origins for amino acids at each residue within each gene family by counting shifts in amino acids with the highest marginal probability at nodes for each site. This number provides an estimate of independent origins of an amino acid at a given site, assuming an equal-rates transition matrix among amino acid substitutions. We acknowledge that this metric will likely be influenced by taxon sampling, as including more species provides greater opportunity to recover convergence (homoplasy). As the

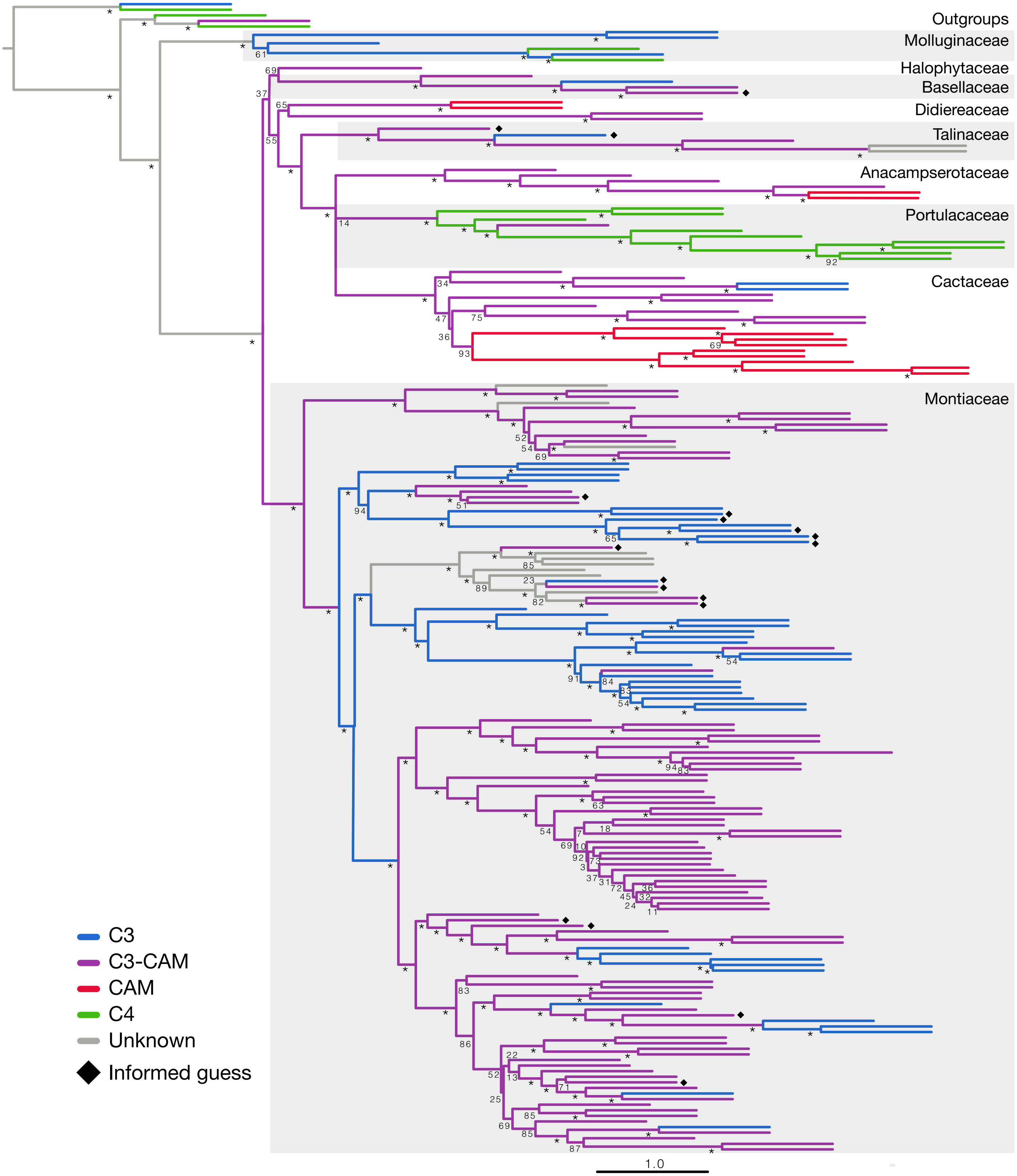

FIGURE 2. Phylogenetic relationships among portullugo species plus outgroups from ASTRAL analysis. ASTRAL bootstrap values are indicated at internal nodes and stars indicate relationships with greater than 95% support. Photosynthetic designations and estimated ancestral states are indicated via color-coded branches (blue = C$_3$, purple = C$_3$-CAM, red = CAM, green = C$_4$, grey = unknown). A diamond indicates a putative pathway assignment that still needs confirmation (see Materials and Methods).

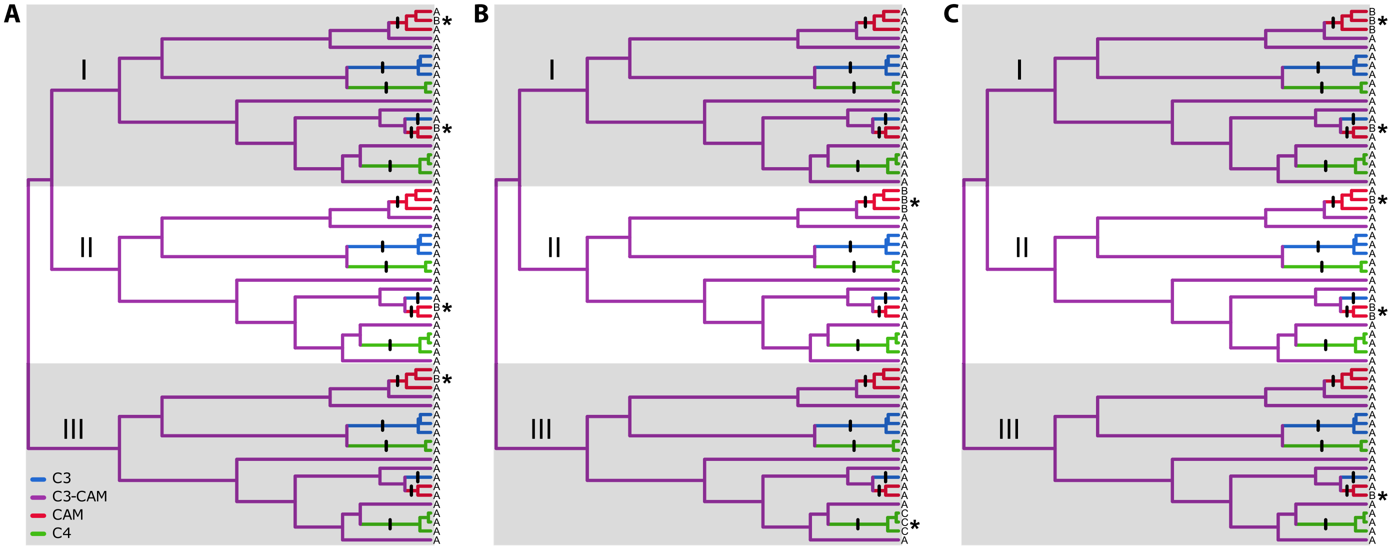

FIGURE 3. Examples of the metrics (*convergent evolution* and *prevalence*) used to identify additional amino acid residues putatively associated with C_3-CAM, CAM, or C_4 pathways. An example gene family with three main paralogs (indicated with roman numerals) is presented, and hypothetical amino acids *A*, *B*, and *C* are given at the tips of each branch, with each panel corresponding to a hypothetical residue. In panel (A), an example of *convergent evolution* of the amino acid *B* is given: the ancestral amino acid is *A*, and *B* evolved in CAM species independently four times (asterisks). In panel (B), an example of amino acid *prevalence* in C_4 and CAM photosynthetic pathways is given: the amino acid *B* is prevalent (present in >50% of species in at least one paralog) in CAM species and rare in C_3 (present in <10% of C_3 species in at least one paralog); likewise, the amino acid *C* is prevalent in C_4 species and rare in C_3 species. In panel (C), the amino acid *B* is both *prevalent* in CAM species and *convergent*, having arisen five times independently.

C_3-CAM phenotype is by far our most common phenotype (103 of 197 tips), we expect that we have biased analyses toward finding convergence in C_3-CAM.

We identified sites in which specific amino acids were prevalent (i.e., occurring in >50% of taxa in at least one ortholog) in either C_3-CAM, CAM, or C_4 species and rare (i.e., occurring in <10% of taxa in at least one ortholog) in C_3 species. Because this analysis is not phylogenetically explicit, the prevalence or rarity of specific amino acids could be confounded (e.g., amplified or diminished) by phylogenetic non-independence. However, the presence of multiple independent origins of both pathways across portullugo could potentially mitigate this problem, and, although both prevalence and convergence may occur by chance (e.g., under relaxed selection), combining both metrics to identify amino acids that are identified as *both* convergent *and* prevalent is our strongest preliminary evidence for molecular adaptation of C_4 or CAM pathways. Due to missing data in gene family alignments, we only considered residues with ≥10% completeness for identifying amino acid prevalence and convergence.

Comparing CCM and non-CCM gene families

To assess whether proteins involved in C_4 and CAM metabolism had higher levels of putatively relevant amino acids than a "background" set of non-CCM genes, we compared our 19 CCM gene families to the remaining non-CCM genes using the number of residues exhibiting evidence for positive selection, the number of instances of convergent evolution, and the number of prevalent C_3-CAM-, CAM-, or C_4-associated amino acids within each gene family. To determine whether gene duplication events might be associated with CCM origins, we compared duplication rates between CCM and non-CCM gene families, as inferred by NOTUNG version 2.8.1.6 (Chen et al., 2000; Stolzer et al., 2012) as implemented in our pipeline (Moore et al., 2017).

RESULTS

Phylogenetic analysis

We recovered 582 loci representing 83 gene families (including 31 subfamilies, e.g., *PPC-1*, *PPC-2*, as designated by Christin et al. 2014, 2015) in 167 species and 197 individuals. Matrix completeness (i.e., the proportion of non-missing nucleotides) for all loci ranged between 26.4% to 100%, with a mean of 67.7% (±16.8 SD). Across loci, average species recovery was 41%, as calculated by the number of species present in a gene tree divided by the number of species descending from the most recent common ancestor in the species tree of the taxa represented in the gene tree. Mean bootstrap support averaged across nodes and loci was 56.8%, with the majority of bootstrap support values concentrated at 0% and 100%. In the coalescent (ASTRAL) species tree (Fig. 2), most major clades within the portullugo were well supported, and relationships are similar to our previous analyses with these loci (Moore et al., 2017), with one key exception. Moore et al. (2017) recovered strong support for Anacampserotaceae + *Portulaca* together as sister to Cactaceae, and in these new analyses, there is instead weak support for a *Portulaca*-Cactaceae clade. Surprisingly, bootstrap support for some of the deeper nodes has decreased with the addition of new taxa here. For example, the ACPT (Anacampserotaceae-Cactaceae-*Portulaca*-*Talinum*) clade is 100% supported, as usual, but the position of Didiereaceae as sister to ACPT (ACPTD) received only modest bootstrap support (55%). Basellaceae and Halophytaceae form a clade, as in Moore et al., but with lowered support, and their position as sister to ACPTD received only 37% bootstrap support. We performed very little curation of this data set, as the species phylogeny was not the primary goal of this project, and predict that removal of the most poorly sampled loci would improve support along the backbone. Because the topology of the tree is roughly consistent with our expectations, we moved forward with our other analyses, which are not based on the species tree but on individual gene family trees. All gene family trees, associated alignments, and the ASTRAL species tree are available in the Dryad Digital Repository (https://doi.org/10.5061/dryad.47m18).

Gene duplications

For most gene families, multiple paralogs were recovered. On average, 6.9 copies (±5.2 SD) were recovered for each gene family, with a range of 1 to 26 paralogs (mean 7.3, ± 6.1 SD, $n = 31$) for the 19 targeted CCM gene families and 1 to 28 paralogs (mean 6.7, ± 4.6 SD, $n = 52$) for the remaining non-CCM gene families. The total number of inferred duplications across genes per branch on the species tree was highly correlated between CCM and non-CCM genes ($R^2 = 0.94$; Appendix S4).

Selection analyses

A mixed effects model of evolution (MEME) was used to identify amino acid residues putatively under positive selection (Murrell et al., 2012). The percentage of sites identified to be under positive selection varied widely across genes, with a mean of 1.7% (±2.3 SD). For CCM gene families, an average of 1.4% (±1.9 SD) of sites was identified, ranging from 0 to 32 sites per gene. For non-CCM gene families, the percentage of sites averaged 1.8% (±2.6 SD), albeit with larger variation in the number of sites across genes, ranging from 0 to 124 sites per gene (Fig. 4). The proportion of sites under positive selection did not significantly differ between CCM and non-CCM gene families (ANOVA: $F_{1,81} = 1.54$, $P = 0.22$; Fig. 4a). Specific sites identified to be under positive selection are displayed in Table 1 and Appendix S5. Amino acid residue numbering for each gene is described in Appendix S3.

Additional MEME selection analyses using gene family trees pruned to C_3-only, C_3-CAM-only, CAM-only, and C_4-only lineages detected, on average, 2.10 (±4.27 SD), 3.86 (±8.89 SD), 0.02 (±0.07 SD), and 0.10 (±0.28 SD) times the number of sites under positive selection as detected by MEME on nonpruned gene family trees, respectively. The percentage overlap in sites under positive selection between pruned and nonpruned analyses for each of the four photosynthetic types was, on average, 11.9% (±26.6 SD), 23.4% (±36.3 SD), 0.57% (±4.1 SD), and 1.1% (±4.9 SD), respectively. However, there was a large discrepancy between CCM and non-CCM gene families: for non-CCM genes, average percentage overlap ranged from 0.0% to 3.8%, whereas CCM genes for trees pruned to C_3-only averaged 32.8% (±36.4 SD) overlap and genes for trees pruned to C_3-CAM-only averaged 59.9% (±36.4 SD) overlap with nonpruned analyses; CAM-only and C_4-only trees for CCM gene families averaged from 1.6% to 3.2%. Amino acid sites identified to be under positive selection for gene family trees pruned to specific photosynthetic types are listed in Appendix S6.

We also performed MEME analyses on individual paralogs based on transcript abundance in C_4 and CAM species (Christin et al., 2015) for the following gene families: aspartate aminotransferase, beta-carbonic anhydrase, NAD malate dehydrogenase, NAD malic enzyme, NADP malate dehydrogenase, NADP malic enzyme, phosphoenolpyruvate carboxylase, phosphate dikinase, and phosphoenolpyruvate-phosphate translocator. Paralogs were divided into putative CCM and non-CCM categories based on transcript abundance (Christin et al., 2015) and run individually in MEME. Results varied widely across gene families for both CCM and non-CCM paralogs: for beta-carbonic anhydrase, NAD malate dehydrogenase, NADP malic enzyme, phosphate dikinase, and phosphoenolpyruvate-phosphate translocator, at least one CCM paralog had more sites under positive selection than non-CCM

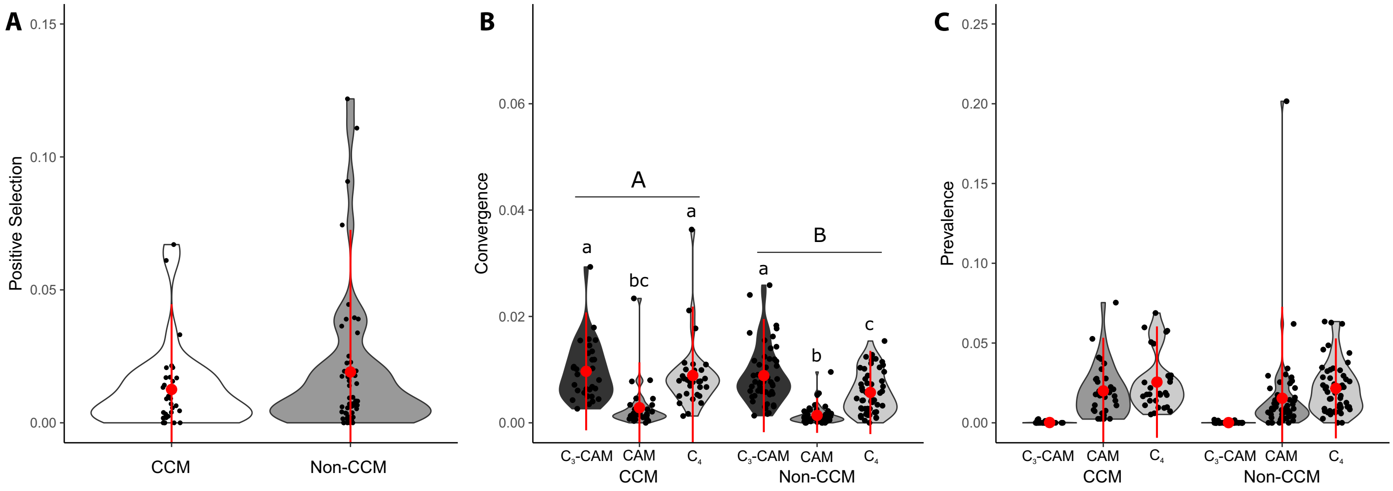

FIGURE 4. Violin plots of (A) the number of sites identified to be under positive selection in carbon concentrating mechanism (CCM) and non-CCM gene families, standardized by sequence length, (B) the total number of independent origins of residues in C_3-CAM, CAM, and C_4 species that lacked origins in C_3 species in CCM and non-CCM gene families, standardized by sequence length and the number of species represented in each pathway, and (C) the number of prevalent sites (see Fig. 3 legend) for C_3-CAM, CAM, and C_4 species in CCM and non-CCM gene families, standardized by sequence length. No significant differences between groups are observed in panels A and C. In panel B, there are a significantly greater number of independent origins for C_3-CAM, CAM, and C_4 species overall in CCM genes than in non-CCM genes (indicated by capital letters), and Tukey groups are assigned with lowercase letters to distinguish significant differences between the interaction of photosynthetic pathway and gene type.

TABLE 1. Sites from carbon concentrating mechanism (CCM) gene families identified by MEME to be under positive selection. For *PPC-1*, parentheses correspond to maize numbering.

Gene family	Sites identified under positive selection ($P \leq 0.05$, false discovery rate)
alaAT	88, 111, 408
AMK-1	None
AMK-2	274, 545
APL-1	29, 332
APL-2	400
APL-3	35, 37, 55, 73, 75, 117, 139, 173, 175, 179, 191, 195, 205, 211, 217, 220, 221, 222, 225, 226, 227, 228, 229, 278, 341, 344, 349, 353, 384, 388, 391, 400
APL-4	399, 472
ASP-1	17, 19, 190, 225, 368, 377
ASP-2	None
ASP-3132	64, 88, 109, 170, 265, 424
BASS	44, 64, 345
betaCA	35, 188, 205, 233, 236, 271, 306
DIC-1112	4, 286, 292
DIC-2	35, 36, 37, 38, 39
DiT-1	248
DiT-2	None
NADMDH	148, 261, 284, 285, 288, 385, 388
NADME-1	11, 311, 471, 558
NADME-2	11, 94, 172, 299, 369, 398, 413, 420, 445, 481, 511, 527, 537
NADPMDH-1	55, 70, 84, 436
NADPMDH-2	67, 199, 308
NADPME	21, 119, 139, 140, 153, 391, 420, 461, 485, 571, 637
NDH	90, 100, 339, 341, 353, 357, 541, 574, 576, 577, 578, 579
PCK	11, 14, 35
PPC-1	475 (480), 548 (553), 569 (574), 603 (608), 650 (655), 823 (828), 834 (839), 864 (869), 868 (873), 873 (878), 874 (879), 892 (897), 899 (904), 905 (910), 906 (911), 907 (912), 909 (914), 911 (916), 913 (918), 917 (922), 918 (923), 919 (924), 920 (925), 921 (926), 922 (927), 923 (928), 929 (934), 930 (930), 934 (938), 935 (939), 937 (941), 954 (958)
PPC-2	None
PPDK	None
PPT-1	246
PPT-2	1, 3, 20, 60, 77, 171, 281
RbcS	6, 8, 15, 65, 67, 81, 91, 102, 103, 137, 170, 178
TDT	151

paralogs, and one CCM NADP malate dehydrogenase paralog had the same number and identity of sites as non-CCM paralogs. For the remaining four of nine gene families, individual paralog analyses recovered fewer sites under positive selection in CCM paralogs than in non-CCM paralogs. Across gene families, CCM paralogs had on average 1.78 (±1.55 SD) times more sites under positive selection, and non-CCM paralogs had on average 2.09 (±3.17 SD) more sites than MEME analyses on nonpruned gene family trees, with a range of 0 to 4.67 and 0 to 12 times the number of sites for CCM and non-CCM paralogs, respectively. Percentage overlap of sites recovered relative to nonpruned analyses was 42.1% (±31.1 SD) and 31.5% (±31.5 SD) for CCM and non-CCM paralogs, respectively. Amino acid sites detected to be under positive selection for selected paralogs are listed in Appendix S7.

Convergence and prevalence analyses

In addition to selection analyses, we searched for molecular convergence (≥2 independent origins) in C_3-CAM, CAM, and C_4 species with zero C_3 origins, as well as prevalence (occurring in ≥50% of taxa) of amino acids in C_3-CAM, CAM, or C_4 taxa with concurrent scarcity (occurring in ≤10% of taxa at least once) in C_3 taxa (Appendices S8 and S9). For both prevalence and convergence assessments, we initially performed analyses with C_3-CAM and CAM pooled into a single group, but were left with an extremely low number of sites detected. Upon reanalysis dividing C_3-CAM and CAM into two separate groups, we were able to identify most of the same sites for prevalence and convergence, as well as several new sites unique to each category, which we present here. Additionally, it is reasonable to think that selection pressures might change when the CAM pathway is used rarely versus when it becomes a primary metabolism. On average, there are 0.009 (±0.005 SD) convergent amino acids per site per C_3-CAM species, 0.002 (±0.003 SD) convergent amino acids per site per CAM species, and 0.007 (±0.005 SD) convergent amino acids per site per C_4 species (ANOVA: $F_{2,\,246} = 52.34$, $p <$ 1e-6). CCM gene families exhibited significantly higher rates of convergent evolution (mean = 0.002, ±0.011 SD) than non-CCM gene families when all photosynthetic physiologies were examined together (ANOVA: $F_{1,\,247} = 6.34$, $P = 0.012$). For CCM and non-CCM genes, respectively, C_3-CAM species exhibited 0.010 (±0.006 SD) and 0.009 (±0.005 SD) convergent amino acids per site per C_3-CAM species, CAM species exhibited 0.003 (±0.004 SD) and 0.001 (±0.002 SD) convergent amino acids per site per CAM species, and C_4 species exhibited 0.009 (±0.006 SD) and 0.0060 (±0.004 SD) convergent amino acids per site per C_4 species (Fig. 4b).

On average, 2.31% (±1.64 SD) and 1.71% (±2.49 SD) of sites were identified to be rare in C_3 species but prevalent in C_4 and CAM species, respectively, whereas only 0.014% (±0.048 SD) of sites were identified as rare in C_3 species but prevalent in C_3-CAM species. Considering CCM genes only, these percentages shifted to 2.55% (±1.75 SD), 2.00% (±1.67 SD), and 0.018% (±0.057 SD) for C_4, CAM, and C_3-CAM species, respectively; while considering only non-CCM genes, these percentages changed to 2.16% (±1.57 SD), 1.54% (±2.87 SD), and 0.012% (±0.042 SD), respectively. The proportion of detected sites did not differ significantly across gene family type (ANOVA: $F_{1,\,81} = 1.40$, $P = 0.24$; Fig. 4C).

DISCUSSION

Lack of elevated positive selection in genes involved in carbon-concentrating mechanisms

Selection tests are used to identify genes that have undergone adaptive evolution. However, it is likely the case that specific regions of protein-coding genes are under very different selective pressures. Popular branch-site and related models extend selection tests by identifying specific protein regions that have likely undergone adaptive evolution (Yang and Reis, 2011; Murrell et al., 2012). In this study, we searched for sites under positive selection in both genes involved in carbon concentrating mechanisms (CCM genes) and non-CCM genes, with the expectation that the extreme variation in photosynthetic pathways would correspond to a detectable difference in CCM genes relative to non-CCM genes. Accordingly, our results appear surprising: in a lineage that harbors multiple transitions between photosynthetic pathways and a variety of unusual photosynthesis phenotypes, we find that genes encoding the major components of C_4 and CAM biochemistry do not appear to be evolving under positive selection to a greater degree than other, presumably random, coding regions of the genome. At first glance, this result seems to stand in sharp contradiction to earlier work that has identified strong selection at multiple sites in key C_4 enzymes (Besnard et al., 2009; Christin et al., 2007, 2014; Rosnow et al., 2014). How could this be? Foremost, there are many ways in which these studies are not directly comparable. Specifically, there are obvious shortcomings to the MEME approach, as well as our data set, though for now we feel like the perceived shortcomings of MEME are in fact what make it such a useful method for us here, due to our limited knowledge of phenotype and gene recruitment.

First, most previous work approached selection tests using methods that a priori identify particular phenotypic states of branches, thus directly investigating how phenotypic states and molecular evolution are coupled (Yang, 1997). MEME does not allow for any phenotypic designations, so we are testing for selection across C_3, C_3-CAM, C_4, and CAM species simultaneously, despite our expectation that each of these photosynthesis types could present very different selection pressures on these proteins. Furthermore, the sampling in this first analysis is skewed heavily toward the Montiaceae, which harbors many C_3 and C_3-CAM plants, and no full CAM or C_4 plants. If we had instead intensely sampled Cactaceae (a highly species-rich lineage of mostly full CAM plants), the analyses might have turned out quite differently.

A second problem concerns the fact that many of the proteins of interest are coded by large gene families, but it is quite possible that only one paralog is involved in C_4/CAM function. Additionally, although it is possible that highly divergent alleles are incorrectly classified as distinct paralogs by our pipeline (or highly similar paralogs may be incorrectly treated as alleles), which could bias tests of positive selection, validation results suggest this potential problem is of relatively minor effect (Moore et al., 2017). Yet we are running analyses across entire gene families simultaneously, and genes not recruited to the new function are likely to be evolving under a very different selection regime—perhaps these are swamping out a signal of positive selection on the subset of genes actually recruited into a new metabolism. This multiple-paralog problem is exacerbated by our lack of knowledge of *which* paralogs are involved in C_4/CAM function, and whether there is strong biased recruitment (sensu Christin et al., 2015) of particular copies across multiple origins. If there is no bias, then distinct paralogs could be C_4/CAM functional in different lineages, making a single strong signal even more difficult to detect.

When we analyzed paralogs separately in the gene families for which we have a reasonable hypothesis about which copies are involved in C_4/CAM function, the results were mixed: some genes showed striking differentiation in detected selection (e.g., *PPDK-1C1a* = 14 sites, all others = 0 sites), while others did not (e.g., NADPME1E1 = 40 sites, all others = 82 sites) (Appendix S7). Again, it is somewhat difficult to determine the reason for these results. Perhaps other NADPME paralogs are active in CAM function in our newly sampled species, and our assumption of biased recruitment is incorrect. Or, perhaps selection has acted on these other paralogs for reasons completely unrelated to C_4 or CAM function.

It is important to note that these complications do not arise from any problem of retrieving the relevant molecular data with hybrid enrichment; rather they stem from our still very limited knowledge about the photosynthetic diversity present in the portullugo, and in particular, how to both identify and delineate relevant photosynthetic phenotypes along a C_3 to full CAM evolutionary trajectory. As we continue to learn more about these species and refine our conceptions of distinct, identifiable, and relevant phenotypes, we can always revisit these analyses.

We are unaware of another study of this size that incorporates MEME analyses, and so want to comment briefly on particular results that concern us. It seems that both the number and identity of the sites that are detected with this approach are *highly* sensitive to taxon sampling, and care should be taken in interpreting results. For example, as discussed above, C_3 and C_3-CAM phenotypes are the most prevalent in our study and should present most of the signal in our data. Yet, when we pruned our gene family trees down to four sets of phenotypically "pure" gene trees, the C_3 and C_3-CAM analyses recovered extremely different numbers of sites (in general, far more), and, what is more disconcerting, the identity of the sites showed little overlap between pruned and nonpruned analyses. We feel that more work is needed to understand the robustness of these inferences to a wide variety of perturbations.

In spite of these caveats, it is worth considering these "negative results" at face value, and how they might inform our nascent understanding of CAM evolution. The most common phenotype in our analyses is what we call C_3-CAM: these species have a functional CAM cycle, but primarily use C_3 photosynthesis, and CAM carbon fixation is secondary—either operating at a very low level at all times, or being upregulated in response to stress, or some combination of the two. We consider our analyses to be most directly a test of positive selection in C_3-CAM species. The lack of elevated selection on C_4/CAM genes in this broad phenotype suggests that any optimization of these proteins occurs later, perhaps as CAM becomes the dominant photosynthetic metabolism. Interestingly, this particular order of events, with enzyme optimization occurring at the latest evolutionary stages, has also been suggested for the assembly of the C_4 syndrome (Christin et al. 2011, 2012; Dunning et al., 2017).

Convergence, prevalence, and parallels between PEPC evolution in portullugo and other lineages

Our convergence and prevalence analyses reveal many putative residues in our investigated gene families that could be important for C_4 and CAM function and indicate significantly higher levels of convergence in CCM than in non-CCM genes (Appendix S8). Both metrics have their caveats: one would expect to discover higher levels of convergence and lower levels of prevalence simply as phylogenetic tree space increases. Both of these trends are visible by comparing the common C_3-CAM phenotype with the poorly sampled CAM phenotype: C_3-CAM demonstrates high levels of convergence yet very low levels of prevalence. It is extraordinary that C_4 phenotypes buck this trend and exhibit the highest levels of convergence, in spite of being represented by only 15 taxa. Furthermore, meaningful comparisons can be made between CCM and non-CCM genes, which on average exhibit similar taxon sampling. Taken together, we feel confident that our analyses have uncovered elevated levels of convergence in CCM genes and that this may represent nascent "optimization" of these genes for C_4/CAM function.

Unfortunately, comparing our identified convergent/prevalent residues to those in other C_4/CAM lineages are mostly limited to genes encoding PEPC, as this is the enzyme that has received the most attention regarding molecular adaptation in C_4 genes. With regard to PEPC, there are notable parallels between patterns of molecular evolution in other groups and C_4 evolution in *Portulaca*. Of the sites identified to be under positive selection in studies of C_4 grasses, sedges, and chenopods (Christin et al., 2007; Besnard et al., 2009; Rosnow et al., 2014), we detected overlap with four sites putatively related to C_4 in our prevalence and convergence analyses. Three of these sites (512, 567, and 568, *Beta* numbering; 517, 572, and 573, maize numbering) exhibited identical amino acid substitutions across C_4 grasses or sedges and most C_4 *Portulaca* (T→A, E→Q, and A→N, respectively). Furthermore, these C_4-associated amino acid substitutions were almost completely confined to *PPC-1E1a'*, which confirms previous results from molecular studies linking *PPC-1E1a'* with the C_4 pathway in *Portulaca* (Christin et al., 2014). Strikingly, these substitutions were absent in the *PPC-1E1a'* copy of *P. cryptopetala*, a known C_3-C_4 intermediate. The fourth site (834/839) was identified to be under positive selection in C_4 sedges and, according to MEME analysis here, in the portullugo. Additionally, this site was identified to be putatively associated with CAM evolution according to prevalence and convergence analyses. Although the transition G→E at this residue is not observed in any investigated C_4 grasses or sedges, nearly 70% of both our C_4 and CAM species exhibited this transition, suggesting a potentially shared molecular adaptation in C_4 and CAM species (via convergent evolution). Our analyses also identified one additional CAM-associated site that was previously detected to be under positive selection in C_4 grasses. However, while C_4 grasses and sedges typically exhibited a D→E transition, the majority of our CAM species (including members of *Anacampseros* and *Alluaudia*) exhibited an E→D transition.

CONCLUSIONS

We have presented molecular analyses of 19 gene families related to C_4 and CAM metabolism across 197 individuals, showing how data can simultaneously inform phylogeny and other aspects of organismal evolution. To our knowledge, no previous study has attempted to integrate functional molecular analyses of large gene families at this scale, nor has previous work compared the results of positive selection analyses against a null expectation using "background" genes (i.e., non-CCM genes here). The lack of a significant difference between the number of sites identified in CCM versus non-CCM genes raises questions about results from exploratory selection analyses in general, as positive selection analyses typically identify multiple sites under selection (Casola and Hahn, 2009; Nozawa et al., 2009; Sironi et al., 2015). Our results suggest that, across a broad sample of genes, it is not unusual to identify some sites as under positive selection, and that detecting selection in hand-picked genes underlying a phenotype of interest may not be documenting what we think. In other words, there may be nothing particularly special about the level of selection on those genes as compared to all genes across the genome. This biased focus is not necessarily a problem if tempered by a comparison to a set of genes that are putatively neutral with regard to the phenotype in question.

Overall, our results highlight consistency with previous results, as well as shared amino acid substitution patterns between C_4 and CAM photosynthetic pathways. Our positive selection results, together with new metrics for identifying sites of interest based on molecular convergence and amino acid prevalence in C_4 and CAM pathways, reveal a large suite of new amino acid substitutions for potential future functional investigation. The ability to use these approaches is a particular advantage of the hybrid enrichment approach to phylogenomics, as we can now efficiently subsample genomes for a handful of target genes with relative ease. Studies of molecular evolution of C_4 and CAM genes have largely been restricted to genes encoding PEPC, and this study expands this sampling considerably, identifying other gene families that appear to exhibit as much potentially significant evolution (e.g., NADME-2, Table 1). Coupled with ecological, physiological, and anatomical work and studies of gene expression, the hybrid enrichment approach to phylogenetics proves to be a powerful tool for developing a truly integrative approach to systematics: resolving species relationships, and subsequently utilizing a robust phylogenetic framework for building a detailed, multi-faceted understanding of organismal evolution.

ACKNOWLEDGEMENTS

The authors sincerely thank Edwards lab members and Chase Mason for comments on the manuscript, Elissa Johnson for illustration assistance, and two anonymous reviewers for suggestions on an earlier draft of the manuscript. This research was funded by the National Science Foundation (DEB-1252901 and DEB-1757427 to E.J.E. and IOS-1708941 to E.W.G.). L.P.H. was supported in part by NSF IGERT grant DGE-0966060. J.M.d.V. was supported in part by Swiss National Science Foundation Fellowship PBZHP3-147199.

DATA ACCESSIBILITY

Voucher information is available in Appendix S1. Raw reads are deposited in the NCBI Short Read Archive (BioProjects PRJNA387599, PRJNA417446, and PRJNA415977). All trees and final alignments are available in the Dryad Digital Repository: https://doi.org/10.5061/dryad.47m18.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

LITERATURE CITED

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. *Journal of Molecular Biology* 215: 403–410.

Benjamini, Y., and Y. Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, B, Methodological* 57: 289–300.

Besnard, G., A. M. Muasya, F. Russier, E. H. Roalson, N. Salamin, and P. A. Christin. 2009. Phylogenomics of C_4 photosynthesis in sedges (Cyperaceae): multiple appearances and genetic convergence. *Molecular Biology and Evolution* 26: 1909–1919.

Casola, C., and M. W. Hahn. 2009. Gene conversion among paralogs results in moderate false detection of positive selection using likelihood methods. *Journal of Molecular Evolution* 68: 679–687.

Chaw, S. M., C. C. Chang, H. L. Chen, and W. H. Li. 2004. Dating the monocot–dicot divergence and the origin of core eudicots using whole chloroplast genomes. *Journal of Molecular Evolution* 58: 424–441.

Chen, K., D. Durand, and M. Farach-Colton. 2000. NOTUNG: a program for dating gene duplications and optimizing gene family trees. *Journal of Computational Biology* 7: 429–447.

Christin, P. A., M. Arakaki, C. P. Osborne, A. Bräutigam, R. F. Sage, J. M. Hibberd, S. Kelly, et al. 2014. Shared origins of a key enzyme during the evolution of C_4 and CAM metabolism. *Journal of Experimental Botany* 65: 3609–3621.

Christin, P. A., M. Arakaki, C. P. Osborne, and E. J. Edwards. 2015. Genetic enablers underlying the clustered evolutionary origins of C_4 photosynthesis in angiosperms. *Molecular Biology and Evolution* 32: 846–858.

Christin, P. A., E. J. Edwards, G. Besnard, S. F. Boxall, R. Gregory, E. A. Kellogg, J. Hartwell, and C. P. Osborne. 2012. Adaptive evolution of C4 photosynthesis through recurrent lateral gene transfer. *Current Biology* 22: 445–449.

Christin, P. A., C. P. Osborne, D. S. Chatelet, J. T. Columbus, G. Besnard, T. R. Hodkinson, L. M. Garrison, et al. 2013. Anatomical enablers and the evolution of C_4 photosynthesis in grasses. *Proceedings of the National Academy of Sciences, USA* 110: 1381–1386.

Christin, P. A., T. Sage, E. J. Edwards, R. M. Ogburn, R. Khoshravish, and R. F. Sage. 2011. Complex evolutionary transitions and the significance of C_3-C_4 intermediate forms of photosynthesis in Molluginaceae. *Evolution* 65: 643–660.

Christin, P. A., N. Salamin, V. Savolainen, M. R. Duvall, and G. Besnard. 2007. C_4 photosynthesis evolved in grasses via parallel adaptive genetic changes. *Current Biology* 17: 1241–1247.

Degnan, J. H., and N. A. Rosenberg. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in Ecology & Evolution* 24: 332–340.

Drummond, A. J., S. Y. Ho, M. J. Phillips, and A. Rambaut. 2006. Relaxed phylogenetics and dating with confidence. *PLoS biology* 4: 88.

Dunn, C. W., A. Hejnol, D. Q. Matus, K. Pang, W. E. Browne, S. A. Smith, E. Seaver, et al. 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature* 452: 745–749.

Dunning, L., J. J. Moreno-Villena, M. R. Lundgren, A. Brautigam, E. J. Edwards, P. Nosil, C. P. Osborne, and P. A. Christin. 2017. Reticulate evolution facilitated the recurrent emergence of C_4 photosynthesis within closely related species. *Evolution* 71: 1541–1555.

Edwards, E. J., and M. J. Donoghue. 2006. *Pereskia* and the origin of the cactus life-form. *American Naturalist* 167: 777–793.

Edwards, E. J., and R. M. Ogburn. 2012. Angiosperm responses to a low-CO_2 world: CAM and C_4 photosynthesis as parallel evolutionary trajectories. *International Journal of Plant Sciences* 173: 724–733.

Faircloth, B. C., J. E. McCormack, N. G. Crawford, M. G. Harvey, R. T. Brumfield, and T. C. Glenn. 2012. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Systematic Biology* 61: 717–726.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution* 17: 368–376.

Felsenstein, J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 39: 783–791.

Gibson, A. C., and P. S. Nobel. 1986. The cactus primer. Harvard University Press, Cambridge, MA, USA.

Grass Phylogeny Working Group II. 2012. New grass phylogeny resolves deep evolutionary relationships and discovers C_4 origins. *New Phytologist* 193: 304–312.

Guralnick, L. J., A. Cline, M. Smith, and R. F. Sage. 2008. Evolutionary physiology: the extent of C_4 and CAM photosynthesis in the genera *Anacampseros* and *Grahamia* of the Portulacaceae. *Journal of Experimental Botany* 59: 1735–1742.

Guralnick, L. J., and M. D. Jackson. 2001. The occurrence and phylogenetics of crassulacean acid metabolism in the Portulacaceae. *International Journal of Plant Sciences* 162: 257–262.

Hillis, D. M. 1987. Molecular versus morphological approaches to systematics. *Annual Review of Ecology and Systematics* 18: 23–42.

Holtum, A. M., L. P. Hancock, E. J. Edwards, and K. Winter. 2017a. Facultative CAM photosynthesis (crassulacean acid metabolism) in four species of *Calandrinia*, ephemeral succulents of arid Australia. *Photosynthesis Research* 134: 17–25.

Holtum, A. M., L. P. Hancock, E. J. Edwards, and K. Winter. 2017b. Optional use of CAM photosynthesis in two C_4 species, *Portulaca cyclophylla* and *Portulaca digyna*. *Journal of Plant Physiology* 214: 91–96.

Jiao, Y., N. J. Wickett, S. Ayyampalayam, A. S. Chanderbali, L. Landherr, P. E. Ralph, L. P. Tomsho, et al. 2011. Ancestral polyploidy in seed plants and angiosperms. *Nature* 473: 97–100.

Jukes, T.H., and C. R. Cantor. 1969. Evolution of protein molecules. *In* H. N. Munro [ed.], Mammalian protein metabolism, 21–132. Academic Press, NY, USA.

Kersey, P. J., J. E. Allen, I. Armean, S. Boddu, B. J. Bolt, D. Carvalho-Silva, M. Christensen, et al. . 2016. Ensembl Genomes 2016: more genomes, more complexity. *Nucleic Acids Research* 44: D574–D580.

Koch, K., and R. A. Kennedy. 1980. Characteristics of crassulacean acid metabolism in the succulent C_4 dicot, *Portulaca oleracea* L. *Plant Physiology* 65: 193–197.

Kosakovsky Pond, S. L., B. Murrell, M. Fourment, S. D. W. Frost, W. Delport, and K. Scheffler. 2011. A random effects branch-site model for detecting episodic diversifying selection. *Molecular Biology and Evolution* 28: 3033–3043.

Kosakovsky Pond, S. L., and S. V. Muse. 2005. Statistical methods in molecular evolution. *In* R. Nielsen [ed.], HyPhy: Hypothesis testing using phylogenies, 125–181. Springer, NY, NY, USA.

Lara, M. V., S. D. Chuong, H. Akhani, C. S. Andreo, and G. E. Edwards. 2006. Species having C_4 single-cell-type photosynthesis in the Chenopodiaceae family evolved a photosynthetic phosphoenolpyruvate carboxylase like that of Kranz-type C_4 species. *Plant Physiology* 142: 673–684.

Lara, M. V., M. F. Drincovich, and C. S. Andreo. 2004. Induction of a crassulacean acid-like metabolism in the C_4 succulent plant, *Portulaca oleracea* L.: study of enzymes involved in carbon fixation and carbohydrate metabolism. *Plant and Cell Physiology* 45: 618–626.

Lemmon, A. R., S. A. Emme, and E. M. Lemmon. 2012. Anchored hybrid enrichment for massively high-throughput phylogenomics. *Systematic Biology* 61: 727–744.

Lewis, P. O. 2001. Phylogenetic systematics turns over a new leaf. *Trends in Ecology & Evolution* 16: 30–37.

Maddison, W. P. 1997. Gene trees in species trees. *Systematic Biology* 46: 523–536.

Mandel, J. R., R. B. Dikow, and V. A. Funk. 2015. Using phylogenomics to resolve mega-families: an example from Compositae. *Journal of Systematics and Evolution* 53: 391–402.

Martin, C. E., and R. S. Wallace. 2000. Photosynthetic pathway variation in leafy members of two subfamilies of the Cactaceae. *International Journal of Plant Sciences* 161: 639–650.

Matasci, N., L. H. Hung, Z. Yan, E. J. Carpenter, N. J. Wickett, S. Mirarab, N. Nguyen, et al. 2014. Data access for the 1,000 Plants (1KP) project. *GigaScience* 3: 1–10.

McCormack, J. E., S. M. Hird, A. J. Zellmer, B. C. Carstens, and R. T. Brumfield. 2013. Applications of next-generation sequencing to phylogeography and phylogenetics. *Molecular Phylogenetics and Evolution* 66: 526–538.

Moore, A. J., J. M. de Vos, L. P. Hancock, E. Goolsby, and E. J. Edwards. 2017. Targeted enrichment of large gene families for phylogenetic inference: phylogeny and molecular evolution of photosynthesis genes in the portullugo clade (Caryophyllales). *Systematic Biology* https://doi.org/10.1093/sysbio/syx078. [Epub ahead of print]

Murrell, B., J. O. Wertheim, S. Moola, T. Weighill, K. Scheffler, and S. L. Kosakovsky Pond. 2012. Detecting individual sites subject to episodic diversifying selection. *PLoS Genetics* 8: e1002764 1–10.

Nevado, B., G. W. Atchison, C. E. Hughes, and D. A. Filatov. 2016. Widespread adaptive evolution during repeated evolutionary radiations in New World lupins. *Nature Communications* 7: article 12384.

Nozawa, M., Y. Suzuki, and M. Nei. 2009. Reliabilities of identifying positive selection by the branch-site and the site-prediction methods. *Proceedings of the National Academy of Sciences, USA* 106: 6700–6705.

Ocampo, G., and J. T. Columbus. 2010. Molecular phylogenetics of suborder Cactineae (Caryophyllales), including insights into photosynthetic diversification and historical biogeography. *American Journal of Botany* 97: 1827–1847.

Ogburn, M. R., and E. J. Edwards. 2013. Repeated origin of three-dimensional leaf venation releases constraints on the evolution of succulence in plants. *Current Biology* 23: 722–726.

Rosnow, J. J., G. E. Edwards, and E. H. Roalson. 2014. Positive selection of Kranz and non-Kranz C_4 phosphoenolpyruvate carboxylase amino acids in Suaedoideae (Chenopodiaceae). *Journal of Experimental Botany* 65: 3595–3607.

Sage, R. F., P. A. Christin, and E. J. Edwards. 2011. The C_4 plant lineages of planet Earth. *Journal of Experimental Botany* 62: 3155–3169.

Sage, R. F., T. L. Sage, and F. Kocacinar. 2012. Photorespiration and the evolution of C_4 photosynthesis. *Annual Review of Plant Biology* 63: 19–47.

Sayyari, E., and S. Mirarab. 2016. Fast coalescent-based computation of local branch support from quartet frequencies. *Molecular Biology and Evolution* 33: 1654–1668.

Schmickl, R., A. Liston, V. Zeisek, K. Oberlander, K. Weitemier, S. C. K. Straub, R. C. Cronn, et al. 2016. Phylogenetic marker development for target enrichment from transcriptome and genome skim data: the pipeline and its application in southern African *Oxalis* (Oxalidaceae). *Molecular Ecology Resources* 16: 1124–1135.

Silvera, K., K. Winter, B. L. Rodriguez, R. L. Albion, and J. C. Cushman. 2014. Multiple isoforms of phosphoenolpyruvate carboxylase in the Orchidaceae (subtribe Oncidiinae): implications for the evolution of crassulacean acid metabolism. *Journal of Experimental Botany* 65: 3623–3636.

Sironi, M., R. Cagliani, D. Forni, and M. Clerici. 2015. Evolutionary insights into host–pathogen interactions from mammalian sequence data. *Nature Reviews Genetics* 16: 224–236.

Stamatakis, A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30: 1312–1313.

Stolzer, M., H. Lai, M. Xu, D. Sathaye, B. Vernot, and D. Durand. 2012. Inferring duplications, losses, transfers and incomplete lineage sorting with nonbinary species trees. *Bioinformatics* 28: i409–i415.

Straub, S. C., M. Parks, K. Weitemier, M. Fishbein, R. C. Cronn, and A. Liston. 2012. Navigating the tip of the genomic iceberg: next-generation sequencing for plant systematics. *American Journal of Botany* 99: 349–364.

Thulin, M., A. J. Moore, H. El-Seedi, A. Larsson, P. A. Christin, and E. J. Edwards. 2016. Phylogeny and generic delimitation in Molluginaceae, new pigment data in Caryophyllales, and the new family Corbichoniaceae. *Taxon* 65: 775–793.

Wickett, N. J., S. Mirarab, N. Nguyen, T. Warnow, E. Carpenter, N. Matasci, S. Ayyampalayam, et al. 2014. Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proceedings of the National Academy of Sciences, USA* 111: E4859–E4868.

Winter, K. 1979. δ13C values of some succulent plants from Madagascar. *Oecologia* 40: 103–112.

Winter, K., and J. A. M. Holtum. 2014. Facultative crassulacean acid metabolism (CAM) plants: powerful tools for unravelling the functional elements of CAM photosynthesis. *Journal of Experimental Botany* 65: 3425–3441.

Winter, K., J. A. M. Holtum, and J. A. C. Smith. 2015. Crassulacean acid metabolism: a continuous or discrete trait? *New Phytologist* 208: 73–78.

Williams, B. P., I. G. Johnston, S. Covshoff, and J. M. Hibberd. 2013. Phenotypic landscape inference reveals multiple evolutionary paths to C_4 photosynthesis. *Elife* 2: e00961.

Yang, Y., M. J. Moore, S. F. Brockington, D. E. Soltis, G. K. S. Wong, E. J. Carpenter, Y. Zhang, et al. 2015. Dissecting molecular evolution in the highly diverse plant clade Caryophyllales using transcriptome sequencing. *Molecular Biology and Evolution* 32: 2001–2014.

Yang, Z. 1996. Among-site rate variation and its impact on phylogenetic analyses. *Trends in Ecology & Evolution* 11: 367–372.

Yang, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Bioinformatics* 13: 555–556.

Yang, Z., and M. dos Reis. 2011. Statistical properties of the branch-site test of positive selection. *Molecular Biology and Evolution* 28: 1217–1228.