Genetic Enablers Underlying the Clustered Evolutionary Origins of C_4 Photosynthesis in Angiosperms

Pascal-Antoine Christin,*[1,2] Mónica Arakaki,[2,3] Colin P. Osborne,[1] and Erika J. Edwards[2]

[1]Department of Animal and Plant Sciences, University of Sheffield, Sheffield, United Kingdom

[2]Department of Ecology and Evolutionary Biology, Brown University

[3]Departamento de Botánica, Facultad de Ciencias Biológicas & Museo de Historia Natural—UNMSM, Lima, Peru

***Corresponding author:** E-mail: p.christin@sheffield.ac.uk.

Associate editor: David Irwin

Abstract

The evolutionary accessibility of novel adaptations varies among lineages, depending in part on the genetic elements present in each group. However, the factors determining the evolutionary potential of closely related genes remain largely unknown. In plants, CO_2-concentrating mechanisms such as C_4 and crassulacean acid metabolism (CAM) photosynthesis have evolved numerous times in distantly related groups of species, and constitute excellent systems to study constraints and enablers of evolution. It has been previously shown for multiple proteins that grasses preferentially co-opted the same gene lineage for C_4 photosynthesis, when multiple copies were present. In this work, we use comparative transcriptomics to show that this bias also exists within Caryophyllales, a distantly related group with multiple C_4 origins. However, the bias is not the same as in grasses and, when all angiosperms are considered jointly, the number of distinct gene lineages co-opted is not smaller than that expected by chance. These results show that most gene lineages present in the common ancestor of monocots and eudicots produced gene descendants that were recruited into C_4 photosynthesis, but that C_4-suitability changed during the diversification of angiosperms. When selective pressures drove C_4 evolution, some copies were preferentially co-opted, probably because they already possessed C_4-like expression patterns. However, the identity of these C_4-suitable genes varies among clades of angiosperms, and C_4 phenotypes in distant angiosperm groups thus represent genuinely independent realizations, based on different genetic precursors.

Key words: **C_4 photosynthesis, crassulacean acid metabolism, transcriptomics, phylogenetics, co-option, evolvability.**

Introduction

During the evolutionary diversification of organisms, novel adaptations emerge through the reassignment of genes inherited from ancestors to novel developmental or biochemical pathways, a process named co-option. The evolutionary accessibility of novel traits can therefore depend on the genomic content of the ancestor (Blount et al. 2008, 2012; Harms and Thornton 2014), together with mutations that produce the new phenotype. But the role of historical contingency in determining the evolutionary potential of specific taxonomic groups remains largely unexplored.

Adaptive traits that independently evolved multiple times represent excellent systems to study the constraints that dictate evolutionary trajectories toward novel adaptations (Fong et al. 2005; Weinreich et al. 2006; Blount et al. 2008, 2012; Marazzi et al. 2012). Among plants, CO_2-concentrating mechanisms (CCMs) rank among the best examples of convergent evolution (Sage et al. 2011). These complex traits consist of numerous anatomical and biochemical components that function together to increase the internal concentration of CO_2 before its fixation by the enzyme ribulose-1,5-bisphosphate carboxylase/oxygenase (Rubisco; Osmond 1978; Hatch 1987), and provide an advantage in warm and arid environments, in the low-CO_2 atmosphere that prevailed for the last 30 My (Sage 2004; Beerling and Royer 2011; Edwards and Ogburn 2012; Sage et al. 2012). They rely on the segregation of the initial fixation of atmospheric CO_2 into organic compounds, which is mediated by the coupled action of carbonic anhydrase (CA) and phosphoenolpyruvate carboxylase (PEPC), and its secondary refixation by Rubisco, which starts the Calvin–Benson cycle (fig. 1). This segregation occurs spatially among distinct compartments within the leaf during the day in C_4 plants, and temporally between night and day in crassulacean acid metabolism (CAM) plants. Besides anatomical requirements, an efficient CCM therefore relies on specific spatial and diurnal expression of CCM-specific genes, as well as suitable catalytic properties of the encoded enzymes (Hibberd and Covshoff 2010; Mallona et al. 2011). Despite their apparent complexity, the C_4 CCM evolved in more than 62 lineages of flowering plants (Sage et al. 2011), and the number of the CAM lineages might be even higher (Edwards and Ogburn 2012). These origins are, however, not randomly distributed in the angiosperm phylogeny, but are clustered within certain clades, whereas other large clades contain no C_4 or CAM species (Sage 2001; Sage et al. 2011; Edwards and Ogburn 2012). This pattern has been attributed to differences in the evolvability of CCMs among angiosperm subclades, because of factors hypothesized to include ecology, life history, and genomic content (Sage 2001; Monson 2003). The association of C_4 origins with particular ecological conditions has since received statistical support (Osborne and Freckleton 2009; Edwards and Smith 2010;

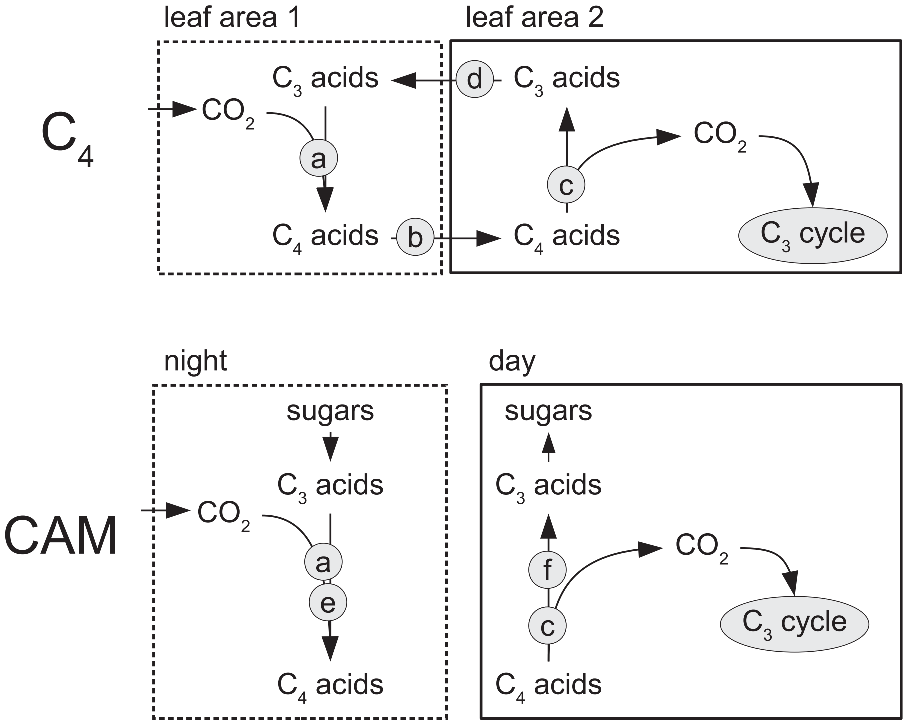

Fig. 1. Simplified schematics of the C_4 and CAM cycles. The main sets of biochemical reactions of the C_4 and CAM cycles are indicated by gray circles, and are separated among the distinct compartments (space or time) that allow separation of initial atmospheric CO_2 fixation (dashed line) and the Calvin cycle (solid line). (*a*) Fixation of atmospheric CO_2 into organic acids through the action of βCA and PEPC, (*b*) transformation and transport of C_4 acids that can involve aspartate aminotransferase (ASP-AT) and malate dehydrogenase (NAD-MDH and/or NADP-MDH), (*c*) decarboxylation of C_4 acids to release CO_2 that can involve malic enzymes (NAD-ME and NADP-ME) or phosphoenolpyruvate carboxykinase (PCK), (*d*) regeneration and transport of C_3 acids by pyruvate, phosphate dikinase (PPDK) and in some cases alanine aminotransferase (ALA-AT), (*e*) transformation of C_4 acids by NADP-MDH and storage in vacuoles, and (*f*) regeneration of C_3 acids by PPDK and storage as starch or sugars.

Kadereit et al. 2012), but the effect of genetic factors is still unknown. The importance of gene duplication for C_4 evolvability has not received empirical support when evaluated through the number of gene copies in complete genomes (Williams et al. 2012). However, the expression profiles and catalytic properties of proteins encoded by a given gene family diversified following speciation events, as well as after gene-specific or whole-genome duplications, and the individual genes present in a given taxonomic group might thus influence the accessibility of the CCMs.

The numerous C_4 origins in the PACMAD subclade of grasses have been statistically associated with the presence of C_4-like anatomical characters in their common ancestor, which were then recurrently co-opted for C_4 evolution, decreasing the number of changes required to generate the C_4 phenotype (Christin, Osborne, et al. 2013; Griffiths et al. 2013). The genetic mechanisms responsible for these anatomical properties are still unknown, but the enzymes responsible for the main reactions of the C_4 and CAM biochemical pathways are well documented (fig. 1; Osmond 1978; Hatch 1987). Their phenotypic variation among taxonomic groups is however not easily quantified, and it is not known whether certain gene lineages encoding these enzymes possess characteristics that facilitate the evolution of the C_4 phenotype. Insights into this question have recently been gained by the comparative analyses of the transcriptomes of three independently evolved C_4 grasses in a phylogenetic context, which showed that only a subset of the genes encoding seven core enzymes were repeatedly co-opted during the evolution of C_4 photosynthesis (Christin, Boxall, et al. 2013), suggesting that C_4-suitable genes were present in the common ancestor of at least some grasses and have been transmitted to most descendants. However, the properties that predisposed certain gene lineages for a CCM function might be more ancient, which could explain the great phylogenetic breadth of origins within angiosperms (fig. 2): perhaps all origins incorporated the same potentiated genes inherited from their common ancestor? This hypothesis can be evaluated by comparing the genetic determinants of CCMs that evolved in distantly related clades of angiosperms.

In this study, we use transcriptome analyses to identify putative CCM-specific gene lineages in different families of Caryophyllales, the clade of angiosperms with the highest recorded number of CCM origins (fig. 2; Sage et al. 2011; Edwards and Ogburn 2012; Kadereit et al. 2012). We then use phylogenetic analyses to identify co-ortholog gene clusters: that is, monophyletic groups of genes that are descended from each of the genes present in the common ancestor of a given clade, through speciation and possibly subsequent gene or genome duplication. These groups of co-orthologs are identified specifically for Caryophyllales, and also for monocots + eudicots. The data available in the

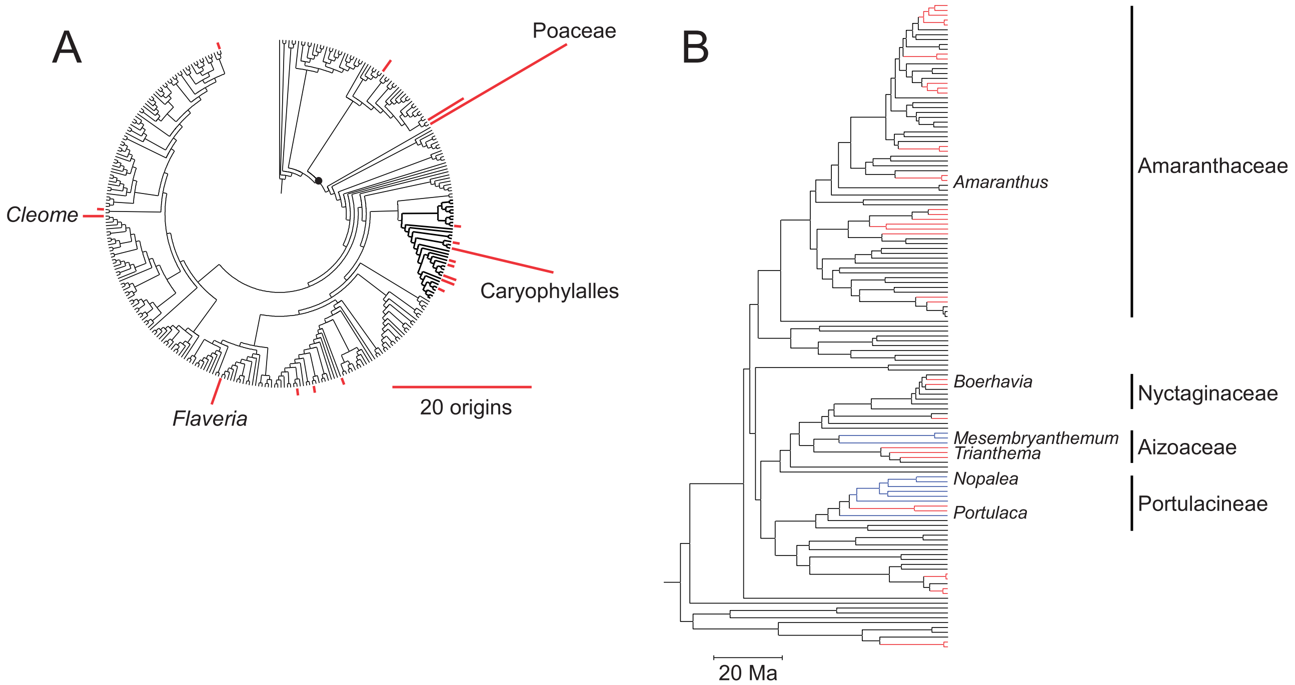

Fig. 2. Phylogenetic position of C_4 lineages. (*A*) Angiosperm phylogeny at the family level inferred by Soltis et al. (2011). The bars on the side are proportional to the estimated number of C_4 origins in each family. The groups discussed in this work are indicated. The split between eudicots and monocots is highlighted by a black circle. Branches within Caryophyllales are in bold. This figure was adapted from Christin and Osborne (2013). (*B*) Time-calibrated phylogeny for Caryophyllales. The tree was inferred from plastid markers (Christin et al. 2011). C_4 taxa are indicated in red, and CAM taxa in blue. The species sampled in this study for transcriptome data are indicated on the right, and the taxonomic groups containing them are delimited with vertical bars.

literature for other C_4 species, and especially C_4 grasses, are also incorporated to test whether 1) there is a bias in gene co-option for CCMs in multiple groups of angiosperms, and 2) the bias is the same for all angiosperms. These analyses shed new light on the importance of historical contingency during evolution and the genetic evolvability of adaptive novelties through time.

Results

Phylogenetic Analyses

Phylogenetic trees were reconstructed for ten gene families encoding enzymes of the C_4/CAM biochemical cycles (fig. 3 and supplementary fig. S1, Supplementary Material online). Between 1 (for genes encoding ALA-AT, NADP-ME, PCK, and PPDK) and 3 (for genes encoding ASP-AT and NAD-MDH), co-ortholog groups across monocots + eudicots were identified. Several of these contain multiple co-ortholog groups specific to eudicots, grasses, or Caryophyllales (fig. 3 and supplementary fig. S1, Supplementary Material online). Relationships among co-orthologs are compatible with the expected species relationships (supplementary fig. S2, Supplementary Material online), despite limited support in some cases and potential problems near the tips due to the presence of tandem repeats that can occasionally recombine (e.g., Wang et al. 2009).

Phylogenetic trees were similarly inferred for ten gene families encoding proteins related to the C_4/CAM traits, but that are not responsible for the core biochemical reactions (supplementary fig. S3, Supplementary Material online). The number of monocots + eudicots co-ortholog groups identified for these families ranged from 1 (*pepck*, *ppdkrp*, *nhd*, and *tdt*) to 5 (*sbas*; supplementary fig. S2, Supplementary Material online).

Expression Patterns of Core C_4/CAM Enzymes

For six of ten gene families encoding core C_4/CAM enzymes (ALA-AT, βCA, NAD-MDH, NADP-ME, PEPC, and PPDK), genes with expression patterns expected for C_4-specific forms were identified for all four C_4 Caryophyllales (*Amaranthus*, *Boerhavia*, *Trianthema*, and *Portulaca*; table 1 and supplementary table S1, Supplementary Material online). The presence of NADP-ME in this list is surprising, given that *Amaranthus* and *Portulaca* use the NAD-ME decarboxylating enzyme (Muhaidat et al. 2007). A high diurnal abundance of *nadpme-1E1* genes could be independent of an involvement in the C_4 cycle, although it is not observed in *Nopalea* and *Mesembryanthemum* (supplementary table S1, Supplementary Material online). Two of the six enzymes with C_4-like expression in the four species are encoded by a single group of co-orthologs in Caryophyllales (ALA-AT and PPDK; supplementary fig. S1, Supplementary Material online). For the four other enzymes, all four species use the same group of co-orthologs in each case (β*ca-2E3*, *nadmdh-3C1*, *nadpme-1E1*, and *ppc-1E1*). An additional β*ca* gene (β*ca-1E2*), not detected at a significant level in the other species, was however present at high levels in *Portulaca*, although it

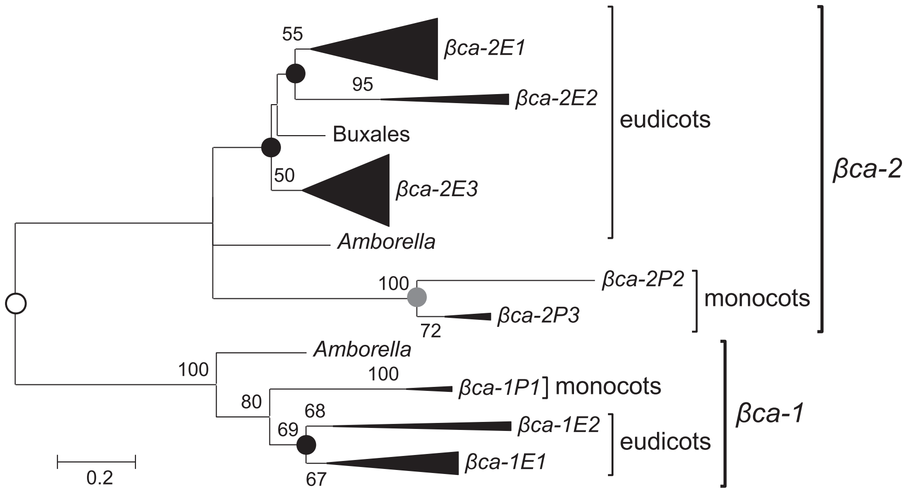

Fig. 3. Example of co-orthologs defined on a gene tree. This tree was inferred for the gene family encoding *β*CA, and bootstrap values are indicated near branches when above 50. Groups of co-orthologs for either eudicots or grasses (=monocots) are compressed, and their name is indicated on the right, as are those of taxonomic groups and co-ortholog groups across monocots and eudicots. Note that *Amborella* is expected to be sister to monocots + eudicots (supplementary fig. S2, Supplementary Material online), and Buxales represents an early branching within eudicots (Soltis et al. 2011). The corresponding gene duplications are indicated by dots, in white if they occurred before the divergence of monocots and eudicots, in gray if they happened in monocots after this divergence, and in black if they happened in eudicots after this divergence. The phylogenetic tree is detailed in supplementary figure S1, Supplementary Material online.

Table 1. Groups of Co-Orthologs Containing Putative C_4- or CAM-Specific Genes in Caryophyllales[a].

Enzyme	C_4				CAM	
	Amaranthus	*Boerhavia*	*Trianthema*	*Portulaca*	*Portulaca*	*Nopalea*
ALA-AT	1	1	1	1	—	—
ASP-AT	3C1	2	—	1E1 + 3C1	—	—
*β*CA	2E3	2E3	2E3	2E3 + 1E1	2E3	2E3
NAD-MDH	3C1	3C1	3C1	3C1a + 2	2	—
NADP-MDH	—	1	1	1	1 + 3	1 + 3
NAD-ME	1	—	—	2	—	—
NADP-ME	1E1	1E1	1E1	1E1a	—	—
PCK	—	—	—	—	—	—
PEPC	1E1	1E1	1E1	1E1a′	1E1c	1E1c + 1E1d
PPDK	1	1	1	1C1b	1C1b	1C1a
AK	—	1	1	1	—	—
PPa	1	1	—	—	—	2
PEPC-K	1	—	—	—	—	—
PPDK-RP	—	—	—	—	—	—
SBAS	1	—	1	1	1	—
DIC	2	—	—	1C2	—	2
DIT	—	—	—	—	—	—
NHD	—	—	1	1	1	—
TDT	—	—	—	—	—	—
PPT	1E2	1E2	—	1E2	1E2	—
TPT	1E2	1E2	1E2	1E2	1E2	—

[a] These were identified based on their transcript abundance. See supplementary table S1, Supplementary Material online, for details of transcript abundance and supplementary figures S1 and S3, Supplementary Material online, for phylogenetic trees and identification of gene lineages.

was still more than 20 times less abundant than *βca-2E3*. An additional *nadmdh* gene (*nadmdh-2*) was similarly present at significant abundance in *Portulaca*, although it was well below the other *nadmdh* gene (supplementary table S1, Supplementary Material online). In addition to these gene families with a significant abundance in all four C_4 species, different lineages encoding ASP-AT have apparently been co-opted in the three C_4 groups with a significant activity (*aspat-1E1*, *aspat-2*, and *aspat-3C1*). Variation in the co-option of genes for ASP-AT was also reported for grasses

(Christin, Boxall, et al. 2013). C_4-like expression was also identified for *nadpmdh-1* in *Boerhavia*, *Trianthema*, and *Portulaca* (table 1 and supplementary table S1, Supplementary Material online). The genes encoding NAD-ME above the rpkm (reads per kilobase per million) threshold were different in *Amaranthus* and *Portulaca*, and in each case, the second gene lineage was just below the threshold (table 1 and supplementary table S1, Supplementary Material online). Proteins encoded by the different *nadme* lineages are known to form heterodimers in *Arabidopsis* (Tronconi et al. 2008), which is also likely the case in these C_4 species. The species *Trianthema portulacastrum* has been described as having a high PCK activity (Muhaidat and McKown 2013), but no *pck* gene was present at high transcript abundance in the samples analyzed here (supplementary table S1, Supplementary Material online). This might indicate that PCK activity varies among *Tr. portulacastrum* individuals or with environmental conditions.

The levels of *ppc-1E1c* genes encoding PEPC increased in the *Portulaca* samples expressing a CAM cycle, as reported in Christin et al. (2014; supplementary table S1, Supplementary Material online). The gene *nadpmdh-3* increased at night, and is likely involved in the CAM pathway of *Portulaca*, together with *nadpmdh-1*, which was present at high levels during both day and night in one of the well-watered samples and reached high levels in all samples expressing a CAM cycle (supplementary table S1, Supplementary Material online). One of the genes encoding NADP-ME (*nadpme-1E1a*) was present at high transcript abundance in *Nopalea* although it did not reach 300 rpkm in both individuals (supplementary table S1, Supplementary Material online). However, putative CAM-specific PPDK and NADP-MDH encoding genes were easily identified during the day period (*ppdk-1C1a* and *nadpmdh-3*; table 1 and supplementary table S1, Supplementary Material online) and another gene encoding NADP-MDH was present at high levels at night (*nadpmdh-1*). Finally, two closely related PEPC-encoding genes were present at high transcript abundance at night (*ppc-1E1c* and *ppc-1E1d*). Although the abundance of *ppc-1E1d* dramatically decreased during the day, the abundance of *ppc-1E1c* remained at similar levels (supplementary table S1, Supplementary Material online).

Expression Patterns of Other Enzymes

One of the genes for adenylate kinase (AK; *ak-1*) was above 300 rpkm during the day in *Boerhavia*, *Trianthema* and *Portulaca*, and one of the genes for inorganic phosphatase (PPa; *ppa-1*) had a high transcript abundance during the day in *Amaranthus* and *Boerhavia*, whereas *ppa-2* was above 300 rpkm during the day in *Nopalea* (supplementary table S1, Supplementary Material online). The single lineage encoding PEPC kinase (PEPC-K) reached high transcript abundance during the day in *Amaranthus* but stayed below 300 rpkm in the other species. However, there was a clear diurnal increase of *pepck-1* in *Trianthema* and *Portulaca*, and a nocturnal increase in *Nopalea* as well as *Portulaca* and *Mesembryanthemum* samples watered less frequently (supplementary table S1, Supplementary Material online). Genes for PPDK regulatory protein (PPDK-RP) stayed at moderate transcript abundance in all samples, with little day/night fluctuations. Regarding the transporters, *sbas-1* was present at high levels in the C_4 *Amaranthus*, *Trianthema* and *Portulaca*, whereas it was at very low abundance in the other species (supplementary table S1, Supplementary Material online), suggesting that it is involved in the C_4 pathway of these three taxa (table 1). Similarly, some genes for dicarboxylate carrier (DIC), sodium:hydrogen antiporter (NHD), and triose-phosphate-phosphate translocator (TPT)/phosphoenolpyruvate-phosphate translocator (PPT) were present at high diurnal abundance in some or all C_4 species (supplementary table S1, Supplementary Material online), supporting their involvement in the C_4 cycle of some species (table 1; Bräutigam et al. 2011, 2014; Külahoglu et al. 2014). The gene *dic-2* for DIC was also present at very high diurnal transcript abundance in *Nopalea* (supplementary table S1, Supplementary Material online), suggesting an involvement in the CAM cycle of this species.

Table 2. Gene Co-Option in Angiosperms[a].

Enzyme	Lineages	Caryophyllales Co-Options[b]	Grass Co-Options[b]	Total Co-Options[b]	Total Co-Opted[c]
ALA-AT	1	4	3	7	1
ASP-AT	3	4	3	7	3
βCA	2	5	3	8	2
NAD-MDH	3	5	0	5	2
NADP-MDH	3	3	3	6	2
NAD-ME	2	2	0	2	2
NADP-ME	1	4	3	7	1
PCK	1	0	2	2	1
PEPC	2	4	3	7	1
PPDK	1	4	3	7	1

[a]Gene co-option in Caryophyllales is based on table 1, whereas gene co-option in grasses is based on Christin, Boxall, et al. (2013; see supplementary table S3, Supplementary Material online). Gene lineages were identified based on phylogenies (supplementary fig. S1, Supplementary Material online), and are listed in supplementary table S1, Supplementary Material online.

[b]Number of times gene lineages were co-opted for C_4 photosynthesis.

[c]Number of different gene lineages that were co-opted at least once for C_4 photosynthesis.

Test for a Bias in Gene Co-Option

Given the size of each gene family, the number of co-ortholog groups in each species, and the number of times each gene family has been co-opted for C_4 photosynthesis (table 2), 18 different gene lineages are expected to be co-opted at least once by chance across seven C_4 origins in monocots and eudicots (fig. 4). The 16 gene lineages identified as C_4-specific across grasses and Caryophyllales (table 2) are not significantly lower than expected by chance ($P = 0.10$). The theoretical minimum is ten genes co-opted at least once (table 2). The observed number is far from this theoretical minimum, and values up to 15 would have been significant (fig. 2), which indicates that the nonsignificance of the test is not simply due to a lack of statistical power.

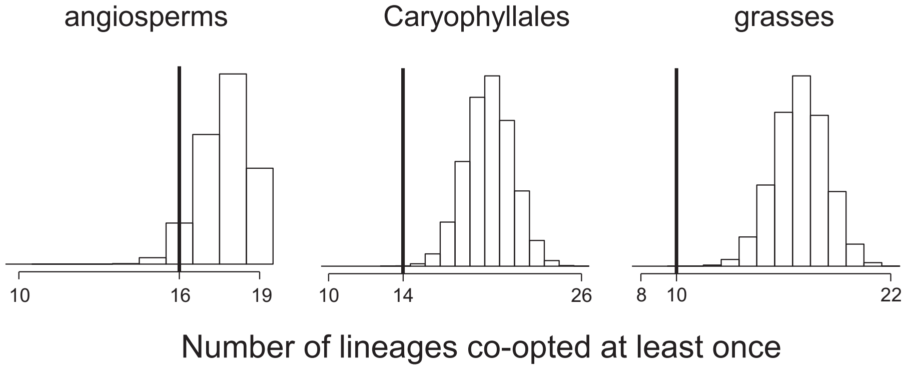

FIG. 4. Test for recruitment bias. The histograms represent the total number of different gene lineages encoding ten enzymes with isoforms involved in the C_4 biochemical pathway, and co-opted for C_4 photosynthesis in 100,000 simulations. The test was conducted across 1) four C_4 origins in Caryophyllales and three in grasses (angiosperms), 2) only the four in Caryophyllales, and 3) only the three in grasses. For each histogram, the vertical bar indicates the observed number of different gene lineages co-opted, and limits of the *x* axis correspond to the minimum and maximum numbers theoretically possible (see table 2 and Supplementary tables S2 and S3, Supplementary Material online; Christin, Boxall, et al. 2013).

Considering only Caryophyllales allows a greater number of co-ortholog groups to be delimited (supplementary table S2, Supplementary Material online). In this clade, an average of 20 different gene lineages that are co-opted at least once during four C_4 origins is expected by chance (fig. 4). The 14 different gene lineages co-opted for C_4 photosynthesis observed in Caryophyllales (supplementary table S2, Supplementary Material online) are significantly lower than expected by chance ($P < 0.0005$), indicating that gene co-option for C_4 photosynthesis was nonrandom, a pattern previously reported within grasses (fig. 4; Christin, Boxall, et al. 2013).

The CAM cycles of *Nopalea* and *Portulaca* use *nadpmdh-3* for part of their cycle, a gene that is not used by any of the sampled C_4 species (supplementary table S1, Supplementary Material online). The co-option of a second *nadpmdh* gene might have been dictated by the need for distinct, diurnally regulated isoforms. Besides this dissimilarity, the other putative CAM-specific genes (*βca-2E3*, *nadpmdh-1*, *ppc-1E1*, and *ppdk-1*) are also used for C_4 photosynthesis in Caryophyllales (table 1), and the co-option of genes for C_4 and CAM photosynthesis can consequently be considered convergent, which might suggest that C_4 and CAM evolution share some genetic predispositions (Christin et al. 2014). Given the size of the four gene families and the co-options observed in C_4 species, the minimum possible total of gene lineages co-opted at least once across the CAM and C_4 origins is 5 (supplementary table S2, Supplementary Material online), which is only one below the total observed from the real data, and the difference is due to *nadpmdh-3*. If the presence of two distinct *nadpmdh* genes is indeed required, then the six observed co-options is the absolute minimum possible number of different co-options given the co-options in C_4. This small number is however not significantly smaller than expected by chance (supplementary fig. S4, Supplementary Material online; $P = 0.22$). There is a good chance that this negative result is due to a lack of statistical power. First, only one origin of CAM was considered. The second CAM sample in our study (*Portulaca*) uses the same gene lineages or very recent duplicates encoding the core CAM enzymes (supplementary table S1, Supplementary Material online), but the CAM pathways of the two groups might not be independent (Christin et al. 2014). In addition, CAM-specific forms were identified for only four enzymes, one of which is encoded by a single-gene family (supplementary table S2, Supplementary Material online), and so the theoretical maximum number of gene lineages that might have been co-opted (10) is also very close to the theoretical minimum (supplementary fig. S4, Supplementary Material online). Finally, the number of subgroups of co-orthologs for some of these is high in *Nopalea* (5 for *ppc-1E1* vs. 1 for *ppc-1E2* and *ppc-2*), making the number of co-orthologs co-opted at least once low even in the absence of a co-option bias (supplementary fig. S4, Supplementary Material online). Testing the hypothesis that CAM origins preferentially co-opted the same genes as C_4 origins in Caryophyllales will require the inclusion of additional independent CAM origins, which exist in Aizoaceae.

Discussion

All Major Angiosperm Gene Lineages Could Be Recruited into C_4 Photosynthesis

An assessment of gene orthology depends strongly on taxonomic scale, as it is affected by clade-specific gene duplications. The divergence between grasses and Caryophyllales (monocots and eudicots) is an early event in the history of angiosperms, which marks the last common ancestor of the majority of species (Soltis et al. 2011). The co-orthologs defined for monocots and eudicots together represent the number of gene copies that existed in their common ancestor, some 130 Ma (Bell et al. 2010; Smith et al. 2010). Multiple groups of deep co-orthologs exist for six of ten enzymes of the C_4/CAM cycles (table 2) and, when considering these, the recruitment of genes for C_4 photosynthesis is not significantly different from random (fig. 4). The gene lineages co-opted by at least one C_4 origin in either grasses or Caryophyllales cover most of the gene lineages present in the common ancestor of monocots and eudicots (table 2 and supplementary table S3, Supplementary Material online). The only exceptions are *ppc*, *nadmdh*, and *nadpmdh*, where one distantly related group of co-orthologs (*ppc-2*, *nadmdh-1*, and *nadpmdh-2*, respectively) was never co-opted, but this is not statistically unexpected given the higher number of gene duplications in the other gene lineages in grasses and Caryophyllales (supplementary fig. S1, Supplementary Material online, and fig. 4). All the other co-ortholog groups for all CCM-related gene families have been co-opted at least once for C_4 photosynthesis, despite sometimes being highly divergent. Most of the gene lineages present in the common ancestor of monocots and eudicots therefore could evolve to encode an enzyme responsible for a C_4 function, at least after tens of million years of further diversification.

. . . But C_4-Potentiation Evolved Later, during the Diversification of Monocots and Eudicots

The different gene families expanded to various degrees after the split of monocots and eudicots through a combination of gene-specific and whole-genome duplications (e.g., Blanc et al. 2003; Flagel and Wendel 2009; Fischer et al. 2014; Rensing 2014), producing up to six distinct co-ortholog groups in certain clades (e.g., *ppc-1* in grasses). For the ten enzymes of the C_4/CAM cycles, the number of gene lineages expanded from 19 to 27 in Caryophyllales, of which only 14 have been co-opted at least once for C_4 photosynthesis (supplementary table S2, Supplementary Material online), which is far smaller than expected by chance (fig. 4). Gene co-option was thus not random, with some genes more likely to become C_4-specific, a pattern also detected within grasses (fig. 4; Christin, Boxall, et al. 2013; John et al. 2014). It has been hypothesized that the presence of closely related copies of one gene following gene duplication favors functional diversification (Zhang 2003), and might have facilitated C_4 evolution (Monson 2003). Although it is certainly true that repeated gene duplications contributed to the functional diversity within gene families, the pattern reported here cannot be simply due to the number of closely related duplicates within each group of co-orthologs, as this was accounted for in our simulations. Other genetic properties must therefore explain the higher C_4-suitability of some of the genes encoding similar enzymes.

As with any complex trait, C_4 photosynthesis must evolve through successive evolutionary steps, each of which has to confer greater fitness than the ancestral one (Heckmann et al. 2013; Williams et al. 2013; Mallmann et al. 2014). As different isoforms of CCM-related enzymes are known to differ in their catalytic properties (Tausta et al. 2002; Svensson et al. 2003; Alvarez et al. 2013), the preferential co-option of genes encoding proteins with C_4-like kinetics might have facilitated C_4 evolution by decreasing the number of mutations required. However, the ability to be directly incorporated into the C_4 cycle might be more important than the kinetic properties of the encoded enzymes. A primitive C_4 cycle might emerge through an increase of PEPC activity sustained by the other enzymes that are already active in the C_3 ancestors (Christin and Osborne 2014). Indeed, some enzymes of the C_4 cycle already exhibit a C_4-like spatial expression in C_3 plants (Hibberd and Quick 2002; Brown et al. 2010; Kajala et al. 2012; Mallmann et al. 2014), which then enables the establishment of a weak C_4 cycle through a few key genetic changes. Once a C_4 pump is acting, natural selection can act to increase its efficiency (Heckmann et al. 2013), through multiple changes to the expression and catalytic properties of each of its constituent enzymes, and the optimization of cellular anatomy. The adaptation of each enzyme through natural selection can however occur only if the enzyme is already involved in the C_4 pathway, making isoforms with C_4-like expression profiles, including cellular and subcellular localization as well as expression levels, more likely to be co-opted. We hypothesize that spatial and temporal expression patterns in the C_3 ancestors affected the likelihood of a given group of co-orthologs being co-opted for CCMs.

Although the exact cause of the co-option bias is not known with confidence, we have clearly established that such a bias exists and that it varies between the two major clusters of C_4 origins in angiosperms. This conclusion is moreover supported by the analysis of other groups of eudicots containing C_4 species (supplementary table S3, Supplementary Material online). Within eudicots, *Cleome* belongs to the Rosidae and *Flaveria* to the Asteridae, whereas Caryophyllales represents an additional lineage (Soltis et al. 2011). These three clades with C_4 species consequently represent ancient splits within eudicots (fig. 2; Bell et al. 2010), and some C_4 enzymes for which the same gene lineage was consistently co-opted within Caryophyllales are encoded by other co-ortholog groups in *Flaveria* (e.g., *ppc-1E2* and *βca-2E1*; supplementary table S3, Supplementary Material online), indicating that the bias evidenced for Caryophyllales does not extend to other eudicots. Homologous transcription factors are apparently involved in C_4 development in monocots and eudicots (Aubry et al. 2014), and comparative studies have established that the mechanism allowing the cell-specificity of some C_4 enzymes was shared between monocots and eudicots, suggesting that they evolved before their split (Brown et al. 2011). However, the functional diversification that happened during the expansion of each gene family has either decreased the C_4-suitability of some gene lineages or produced gene lineages that are more suitable for a C_4 function, through modifications that predate C_4 photosynthesis and consequently evolved for unrelated reasons. The modifications that favored co-option for C_4 photosynthesis are probably different for each gene family, and likely happened at different times. Interestingly, the same gene lineages are used for the C_4 and CAM pathways in the closely related species included in this study (table 1). Although more CAM origins are needed to confirm this hypothesis, it might indicate that genes more likely to be co-opted for C_4 are equally suitable for CAM photosynthesis, which might result from shared requirements for high expression levels in photosynthetic organs (for some enzymes in the same cells and same time of day; e.g., PPDK and MDH) and kinetic properties adapted to high concentrations of substrates and products.

Conclusions

The repeated evolution of CO_2-concentration mechanisms in flowering plants represents one of the best examples of convergent adaptation to changing environments, and constitutes an extraordinary system to test hypotheses about the effect of ancestral states on the evolutionary trajectories of descendants. It has been shown that C_4 origins in grasses were constrained to subclades with anatomical enablers (Christin, Osborne, et al. 2013), and the observation of a gene co-option bias within the same group suggested that genetic enablers might also have contributed to C_4 evolvability (Christin, Boxall, et al. 2013). By expanding the sampling to distantly related lineages that also contain multiple C_4 origins, we have shown here that the gene co-option bias is not restricted to grasses, but is also present in the distantly related Caryophyllales, where it might even extend to CAM origins. However, although both clades show such a bias, it is not present when they are analyzed jointly, which indicates that the bias is clade-specific. The changes that increased the suitability of some genes for the C_4 function happened during the diversification of angiosperms and not before the divergence of eudicots and monocots, nor before the divergence of the major groups of eudicots (e.g., Rosidae, Asteridae and Caryophyllales). Therefore, the C_4 origins found in tight phylogenetic clusters might be best considered as parallel realizations of a C_4 syndrome, which have repeatedly evolved from a potentiated state that was inherited from their C_3 ancestor. The C_4 origins observed in Caryophyllales are however clearly independent from those observed in grasses. This is likely also true of the other, smaller clusters of C_4 origins present in angiosperms (fig. 3; Sage et al. 2011), revealing the complex contributions of parallelism and convergence throughout the evolutionary history of C_4 photosynthesis.

Materials and Methods

Plant Material

A total of six species of the Caryophyllales were selected for quantitative transcriptome analyses. The species *Portulaca oleracea* is constitutively C_4 but a complementary CAM pathway can be triggered by drought stress (Kraybill and Martin 1996; Lara et al. 2003; Winter and Holtum 2014). The transcriptome of *P. oleracea* individuals expressing the C_4 and CAM cycles, respectively, was sequenced in a previous study, including analyses of selected genes (Christin et al. 2014). The additional species in this study include three C_4 taxa (*Amaranthus hypochondriacus*, *Boerhavia coccinea*, and *Trianthema portulacastrum*), one constitutive CAM species (*Nopalea cochenillifera*), and one CAM-inducible plant (*Mesembryanthemum crystallinum*) from distinct Caryophyllales families. These different species are separated in the phylogeny by multiple C_3 taxa and are thought to represent different origins of the C_4 and CAM pathways (fig. 2; Arakaki et al. 2011; Brockington et al. 2011; Sage et al. 2011), although the CAM cycles of *N. cochenillifera* and *P. oleracea* might share a partially common origin (Christin et al. 2014).

The plants were grown from seeds, except for *N. cochenillifera* samples, which were acquired as small plants. All seedlings were placed simultaneously in a Conviron E7/2 plant growth chamber (Conviron Ltd., Winnipeg, MB, Canada). The conditions were as described in Christin et al. (2014), with 14 h of light, and a night temperature of 22 °C, which increased to 28 °C after 3 h of light and until 3 h before dark. The chamber was illuminated with twelve 32-W fluorescent lamps and four 60-W incandescent lamps. Each plant was grown individually in a 7.5-cm pot, except for *N. cochenillifera* individuals, which were bigger and had to be placed in 445-ml pots. The pots were filled with cleaned mix for succulent plants (2 parts soil, 1 part perlite, 1 part gravel, and 1 part calcined clay). Their position within the growth chamber was randomized daily for the duration of the experiment. Most plants were bottom-watered as needed to keep the soil constantly moist. However, one group of *P. oleracea* and *M. crystallinum* seedlings were selected at the beginning of the experiment and were watered less frequently to induce a CAM cycle. Nutrients were added to the water periodically at a concentration of 1:100 (w/v) of K, P, and N in equal proportions.

After 1 month in these conditions, leaf samples (or stem fragments for *N. cochenillifera*) were collected and flash-frozen in liquid nitrogen and stored at −80 °C. For each species, two individuals were sampled after 4 h of light (day sample) and after 2 h of dark (night sample). For *P. oleracea* and *M. crystallinum*, two individuals of each watering regime were sampled. For each species, one individual was sampled first during the day and then on the consecutive night, and the other one was sampled first at night and then during the consecutive day. This sampling was meant to control for effects triggered by the removal of leaves. An equal proportion of young and mature leaves were sampled. Multiple leaves from each sample were randomly mixed for RNA extraction.

In addition to these 32 samples used for quantitative transcriptome analyses, eight individuals were sampled for qualitative transcriptome analyses, without controlling for the growth conditions or tissue type. These additional individuals represent different families of the suborder Portulacineae (*Talinum portulacifolium*, *Anacampseros filamentosa*, *Pereskia grandifolia*, *Pe. bleo*, *Pe. lychnidiflora*, *Echinocereus pectinatus*, and another individual of both *P. oleracea* and *N. cochenillifera*). The sampled individuals came from the collection of living material available in the greenhouse of Brown University, except for *P. oleracea*, which was collected in Providence, RI. For each individual, various proportions of leaf, stem, root, and/or floral tissue were mixed before RNA extraction to increase the diversity of transcripts in the sequencing results.

RNA Extraction, Sequencing, and Assembly

RNA isolation was performed using the RNeasy Plant Mini Kit (Qiagen Inc., TX), or for the succulent tissues the FastRNA Pro Green Kit (MP Biomedicals US, OH), and included a DNAse treatment. For each sample (one individual in one condition), several extractions were performed and pooled. The samples were prepared for sequencing using the Illumina TruSeq mRNA Sample Prep Kit (Illumina Inc., CA) and following the provider's instructions. Each sample was tagged with a specific barcode. Fragments of the cDNA libraries ranging from 400 to 450 bp were selected and sequenced as pairedend 100-bp reads using the Illumina HiSeq 2000 instrument at Brown University Genomics Core Facility. For the quantitative transcriptome analyses, 16 samples were pooled per lane, whereas the samples for the qualitative transcriptome analyses were sequenced on one-ninth of a lane. Raw reads were deposited in NCBI SRA database, under the project accession SRP050968. Accession numbers for individual samples are indicated in supplementary table S4, Supplementary Material online.

The reads from each sample were assembled individually to decrease assembly difficulties caused by different alleles among individuals. The assemblies were performed using the software Trinity (Grabherr et al. 2011) as implemented in the Agalma pipeline (Dunn et al. 2013; supplementary table S4, Supplementary Material online). For the quantitative transcriptome samples, reads for each individual were mapped to the assembled contigs using the software Bowtie 2 (Langmead and Salzberg 2012; supplementary table S4, Supplementary Material online), which is a reliable method to estimate transcript abundance (Marioni et al. 2008; Siebert et al. 2011). The mixed model was used, which allows unpaired alignments when paired alignments fail. Only one of the best alignments was reported per read, and the number of reads mapping each contig was used to compute reads per million or reads (rpm), which were later transformed into rpkm (see below). Multiple best alignments were frequent because multiple contigs were generally assembled per locus. These contigs were however merged during the phylogenetic annotation (see below), so that the rpm value per gene was not affected.

Reference Data Sets and Phylogenetic Trees

The assembled contigs corresponding to genes encoding proteins that are involved in the C_4 or CAM cycles were identified and annotated phylogenetically using an improved version of the approach developed by Christin, Boxall, et al.(2013) and Christin et al. (2014). A list of C_4- and CAM-related proteins was compiled from the literature (Osmond 1978; Hatch 1987; Bräutigam et al. 2011, 2014; Christin, Boxall, et al. 2013). This list is not clade-specific, nor relevant for a specific biochemical subtype only. It includes ten core enzymes, which were used to test a bias in gene co-option (see below). In addition, ten gene families potentially related to CCMs were analyzed as they might be important to engineer CCMs in C_3 crops (Bräutigam et al. 2014). This includes two enzymes involved in the processing of PPDK products, two regulatory proteins, as well as six metabolite transporters (one of the gene families includes genes encoding PPT and others encoding TPT; supplementary fig. S3 and table S1, Supplementary Material online). The annotation was performed in two consecutive steps. First, a reference data set was compiled for each enzyme from sequences extracted from complete genomes, public databases, and the longest of the contigs assembled here. This was used to infer high-quality phylogenetic trees and identify co-ortholog groups. These groups are defined as all the genes descending from a given speciation event, and are consequently specific for a given taxonomic group. Then, all the homologous contigs assembled here were successively compared with the corresponding reference data set, and assigned to one of the co-ortholog groups.

For the reference data sets, coding sequences of all genes encoding each of the selected enzymes were retrieved from *Arabidopsis thaliana* based on their annotation. *Arabidopsis* was used because it has the best annotation, but any species could have been used as the starting genome. In some cases, well-annotated sequences for other species were retrieved from GenBank and added to this reference data set. These sequences were used as the query of a Basic Local Alignment Search Tool (BLAST) search against *Arabidopsis*-predicted cDNAs based on its genome with a maximal *e* value of 0.0001. The identified homologous sequences were added to the reference data set, which was recursively used as the query of a BLAST search against additional predicted cDNAs based on 17 complete genomes (supplementary fig. S2, Supplementary Material online), each time adding the homologous sequences to the reference data set. Each data set was manually curated and sequences that were similar on a small fragment only or that were obviously incomplete or corresponded to chimeras were removed. The data set for each gene family was completed with sequences extracted from the transcriptomes generated in this study, using the same approach except that only contigs that matched the reference sequences on at least two-thirds of the average length of the other coding sequences were retained, and an *e* value of 0.01 was used. These new reference data sets were aligned and manually inspected. The congruence between the sequences assembled here from Illumina data and those generated by other sequencing methods and extracted from public databases confirmed the quality of our assemblies. Putative introns and untranslated regions were identified based on homology, the GT–AG rule, and start and stop codons, and these noncoding regions were removed. All contigs that had indels affecting the reading frame were deleted. The remaining sequences were translated into amino acid sequences and aligned using ClustalW (Thompson et al. 1994), and a phylogenetic tree was inferred on the nucleotide sequences using Phyml (Guindon and Gascuel 2003). Groups of very similar contigs from the same species were identified and only the longest sequence of each was retained. The selected sequences were again visually inspected, and possible chimeras between closely related paralogs were identified using the software Geneconv (Sawyer 1999) and were removed. The remaining sequences extracted from either the complete genomes or transcriptomes generated here constituted the reference data sets used for further phylogenetic annotation.

For each gene family, the alignment of the reference dataset was manually refined, and extremities were truncated to remove regions that were too variable to be unambiguously aligned. A phylogenetic tree was then computed on nucleotide sequences using Phyml and the best-fit substitution model identified through hierarchical ratio tests (GTR + G or GTR + G + I in all cases), with 100 bootstrap pseudoreplicates. After this phylogenetic tree revealed two very distant groups in some families, trees were computed separately for each of these groups. Each phylogenetic tree was manually inspected and co-ortholog groups common to grasses and eudicots were identified, as monophyletic groups of genes congruent with the species relationships (fig. 3). In some cases, one gene lineage identified for grasses and eudicots contained multiple co-ortholog groups in either eudicots or grasses, and these were annotated as such (fig. 3). Similarly, some groups of eudicot co-orthologs contained several co-ortholog groups in Caryophyllales and some groups of Caryophyllales co-orthologs contained several co-ortholog groups in Portulacineae. These different levels of orthology were all considered, so that gene lineages were defined for Caryophyllales and Portulacineae and could be matched to a more inclusive set of eudicot co-orthologs for comparison with distantly related clades, and to a most inclusive set of monocots + eudicots co-orthologs for comparison with grasses. Numbers attached to the gene names were used to describe these groups of monocots + eudicots co-orthologs. Grass-specific co-ortholog groups were named by adding to the angiosperm name a “P” (for Poaceae) and the number that was given previously (Christin, Boxall, et al. 2013). In cases where co-ortholog groups specific to eudicots were detected, an “E” (for eudicots) was added to the monocots + eudicots name, and these were numbered consecutively. In some cases, lack of phylogenetic support prevented the identification of eudicot-specific co-ortholog groups, but these could be easily identified for Caryophyllales (e.g., *aspat-3*; see Results), and were consequently named using the same rules with a “C” instead of “E.” Finally, in some cases, distinct co-ortholog groups specific to the Portulacineae subclade of Caryophyllales were observed and named adding lower-case letters (e.g., *ppdk-1C1a* and *ppdk-1C1b*) to the Caryophyllales-specific name. The only exception is *ppc-1E1* where the numerous Portulacineae-specific co-ortholog groups have been identified with a dense species sampling and genomic DNA and named by Christin et al. (2014). The identity of the C_4-specific genes for two additional lineages of C_4 eudicots (*Cleome*, Brassicales and *Flaveria*, Asterales) was retrieved from the literature (supplementary table S3, Supplementary Material online; Borsche and Westhoff 1990; Rosche and Westhoff 1990; McGonigle and Nelson 1995; Tanz et al. 2009; Bräutigam et al. 2011; Gowik et al. 2011). These genes were assigned to co-ortholog groups identified within the phylogenetic trees inferred in this study, based on the presence of either *Flaveria* genes or *Arabidopsis* genes orthologous to *Cleome* in the trees.

Phylogenetic Annotation of All Contigs and Transcript Abundance

Once the reference data sets and corresponding phylogenetic trees were available, all contigs from the 32 quantitative transcriptomes belonging to each gene family were identified through BLAST searches, with the reference data set used as the query against each transcriptome, and an *e* value of 0.01. Each of the identified contigs was then individually placed in the phylogenetic tree. The matching region of the contig, identified through the BLAST search, was added to the reference data set, which was aligned with Muscle (Edgar 2004), a program that can be easily automated, and a phylogenetic tree per contig was inferred with Phyml and a GTR + G model. The phylogenetic tree was automatically inspected, and the contig was assigned to a group of Caryophyllales or Portulacineae co-orthologs identified based on the reference data set if they formed a monophyletic group. In order to differentiate the Portulacineae-specific *ppc-1E1* copies, identified with a large sample of sequences isolated from genomic DNAs (Christin et al. 2014), the *ppc-1E1* transcripts from *Nopalea* and *Portulaca* were reannotated using the same method but with a reference data set comprised of a sample of Portulacineae sequences previously isolated that represented the different gene lineages.

The rpm values for all the contigs assigned to a given group of co-orthologs were summed to obtain the rpm value for each gene lineage. These rpm values were then transformed into rpkm values, based on the length of orthologous mRNAs for model organisms. The length of the assembled contigs was not used because they do not generally cover the whole length of the transcript, and rpkm values for contigs covering parts of the same transcript cannot be added. The rpkm values were used to identify the groups containing C_4- and CAM-specific genes, which are routinely identified based on quantitative gene expression with consistent results among studies (Bräutigam et al. 2011, 2014; Gowik et al. 2011; Mallmann et al. 2014). For each species, genes were considered as putative C_4 forms if they were present at more than 300 rpkm during the day in both replicates. The same criterion was used to identify putative CAM forms, except that a predominantly nocturnal expression was expected for several of them (fig. 1). To examine whether our 300 rpkm cutoff influenced our results, we reran the test for biased co-option in Caryophyllales (see below) with an extremely conservative cutoff of 1,000 rpkm and, although fewer gene lineages were identified as putatively C_4-specific, the test was still significant, as only one group of co-orthologs per multigene family contains gene lineages with rpkm values above 1,000 (see supplementary table S1, Supplementary Material online).

Tests for C_4 Recruitment Bias

The ten gene families encoding enzymes responsible for the main C_4 biochemical reactions were considered to test the hypothesis that the co-option of particular groups of co-orthologs within each gene family was not random. It is well established that the activity and transcript abundance of the C_4-specific forms of these gene families differ from those observed in C_3 species (Bräutigam et al. 2011, 2014; Christin, Boxall, et al. 2013; Külahoglu et al. 2014; Mallmann et al. 2014), so that a high transcript abundance in a C_4 group can be considered as an evolutionary novelty and is consequently independent from a high transcript abundance in another C_4 group. Gene families encoding enzymes linked to the processing of PPDK products, regulation of C_4 enzymes, and transport of metabolites were not included in the test because their expression level in C_3 species can also be high (Christin, Boxall, et al. 2013) and is generally not known with confidence, so that a high transcript abundance in different C_4 lineages could be inherited from their common ancestor and would consequently be nonindependent.

A bias in gene co-option for C_4 photosynthesis was first tested across monocots + eudicots. For each of the ten gene families, the number of co-ortholog groups across both monocots + eudicots was determined using phylogenetic trees (supplementary fig. S1, Supplementary Material online; table 2). The transcriptome data generated in this study were used to estimate the number of co-option events in Caryophyllales, which can exceed the number of C_4 groups if more than one gene lineage is used by the same species (e.g., genes encoding βCA and NAD-MDH in *Portulaca*). The number of different gene lineages co-opted at least once was also recorded. The same information was retrieved for three independent origins of C_4 in grasses, using the analyses of Christin, Boxall, et al. (2013). We then generated the null distribution that would be expected if co-option of gene lineages was unbiased, by simulating a random recruitment from the total pool of available lineages across the ten gene families, using 100,000 replicates. These simulations accounted for the number of co-ortholog groups in each family as well as the number of gene duplications in each lineage, estimated by the number of clade-specific subgroups of co-orthologs. For instance, the gene *aspat-1* encoding ASP-AT is present as two co-ortholog groups in grasses (supplementary fig. S1, Supplementary Material online), and is therefore twice as likely to be co-opted by chance in this group than *aspat-2* and *aspat-3*. The *P* value associated with the hypothesis of co-option bias was computed as the number of replicates producing a total of gene lineages co-opted that were fewer or equal to the number observed in the real data set.

The hypothesis of a bias in gene co-option for C_4 photosynthesis was then tested specifically for Caryophyllales. The number of gene lineages differed from the previous analysis because some co-ortholog groups defined for monocots + eudicots contain multiple co-ortholog groups in Caryophyllales (supplementary fig. S1, Supplementary Material online). Again, the number of gene lineages, as well as multiple subgroups of co-orthologs in some taxa (e.g., *nadmdh-3* in Portulacineae), was accounted for. The exact same test was also applied to the data set from grasses (Christin, Boxall, et al. 2013).

Test for CAM Recruitment Bias

A similar approach was used to test the hypothesis that CAM origins preferentially co-opt genes used by C_4 groups. This test was conducted using Caryophyllales only, and C_4 gene co-option was fixed to that observed in the real data. The hypothesis tested was that the total number of gene lineages co-opted at least once across the four C_4 and the CAM origins does not differ from that expected if the co-option for CAM was independent of the identity of genes co-opted for C_4. The total number of lineages co-opted at least once across the C_4 or CAM origins expected by chance was obtained by randomly selecting genes for CAM photosynthesis, and adding them to the tally of gene lineages co-opted in the four C_4 origins. This test was based on the four enzymes for which CAM-specific isoforms were identified (βCA, NADP-MDH, PEPC, and PPDK; supplementary table S2, Supplementary Material online), and the presence of multiple subgroups of co-orthologs in some species was again accounted for. Only *Nopalea* was considered, as the CAM pathway of *Portulaca* might not represent a completely independent origin (Christin et al. 2014), and our *Mesembryanthemum* transcriptome data did not strongly support a successful induction of a CAM cycle (supplementary table S1, Supplementary Material online). Indeed, the samples of *Mesembryanthemum* grown with different watering regimes did not differ markedly in their transcript abundances (supplementary table S1, Supplementary Material online). The duration of the drought period was probably too short to trigger a high-level CAM cycle (Winter and Holtum 2014) and these samples were consequently not used to detect CAM-specific genes. However, two *nadpmdh* genes (*nadpmdh-1* and *nadpmdh-3*), *ppc-1E1*, and *βca-2E3* increased in nocturnal abundance in the low water treatment, and *nadpme-1* increased during the day, which might suggest that this independent CAM origin uses the same isoforms as *Portulaca* and *Nopalea* (see Results; supplementary table S1, Supplementary Material online).

Supplementary Material

Supplementary tables S1–S4 and figures S1–S4 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

This work was funded by a Marie Curie International Outgoing Fellowship (grant number 252569) and a Royal Society Research Fellowship (grant number URF120119) to P.A.C., and a National Science Foundation grant (grant number DEB-1026611) and Brown University Salomon Faculty Research Award to E.J.E. This research was conducted using computational resources at the Center for Computation and Visualization, Brown University. The authors thank Professor Alan Lloyd, who provided seeds for *Boerhavia coccinea*.

References

Alvarez CE, Saigo M, Margarit E, Andreo CS, Drincovich MF. 2013. Kinetics and functional diversity among the five members of the NADP-malic enzyme family from *Zea mays*, a C^4 species. *Photosyn Res.* 115:65–80.

Arakaki M, Christin PA, Nyffeler R, Lendel A, Eggli U, Ogburn RM, Spriggs E, Moore MJ, Edwards EJ. 2011. Contemporaneous and recent radiations of the world's major succulent plant lineages. *Proc Natl Acad Sci U S A.* 108:8379–8384.

Aubry S, Kelly S, Kümpers BM, Smith-Unna RD, Hibberd JM. 2014. Deep evolutionary comparison of gene expression identifies parallel recruitment of trans-factors in two independent origins of C_4 photosynthesis. *PLoS Genet.* 10:e1004365.

Beerling DJ, Royer DL. 2011. Convergent Cenozoic CO_2 history. *Nat Geosci.* 4:418–420.

Bell CD, Soltis DE, Soltis PS. 2010. The age and diversification of the angiosperms re-revisited. *Am J Bot.* 97:1296–1303.

Blanc G, Hokamp K, Wolfe KH. 2003. A recent polyploidy superimposed on older large-scale duplications in the *Arabidopsis* genome. *Genome Res.* 13:137–144.

Blount ZD, Barrick JE, Davidson CJ, Lenski RE. 2012. Genomic analysis of a key innovation in an experimental *Escherichia coli* population. *Nature* 489:13–518.

Blount ZD, Borland CZ, Lenski RE. 2008. Historical contingency and the evolution of a key innovation in an experimental population of *Escherichia coli*. *Proc Natl Acad Sci U S A.* 105:7899–7906.

Borsch D, Westhoff P. 1990. Primary structure of NADP-dependent malic enzyme in the dicotyledonous C_4 plant *Flaveria trinervia*. *FEBS Lett.* 273:111–115.

Bräutigam A, Kajala K, Wullenweber J, Sommer M, Gagneul D, Weber KL, Carr KM, Gowik U, Mass J, Gowik U, et al. 2011. An mRNA blueprint for C_4 photosynthesis derived from comparative transcriptomics of closely related C_3 and C_4 species. *Plant Physiol.* 155: 142–156.

Bräutigam A, Schliesky S, Kulahoglu C, Osborne CP, Weber APM. 2014. Towards an integrative model of C_4 photosynthetic subtypes: insights from comparative transcriptome analysis of NAD-ME, NADP-ME, and PEP-CK C_4 species. *J Exp Bot.* 65:3579–3593.

Brockington SF, Walker RH, Glover BJ, Soltis PS, Soltis DE. 2011. Complex pigment evolution in the Caryophyllales. *New Phytol.* 190:854–864.

Brown NJ, Newell CA, Stanley S, Chen JE, Perrin AJ, Kajala K, Hibberd JM. 2011. Independent and parallel recruitment of preexisting mechanisms underlying C_4 photosynthesis. *Science* 331:1436–1439.

Brown NJ, Palmer BG, Stanley S, Hajaji H, Janacek SH, Astley HM, Parsley K, Kajala K, Quick WP, Trenkamp S, et al. 2010. C_4 acid decarboxylases required for C_4 photosynthesis are active in the mid-vein of the C_3 species *Arabidopsis thaliana*, and are important in sugar and amino acid metabolism. *Plant J.* 61:122–133.

Christin PA, Arakaki M, Osborne CP, Bräutigam A, Sage RF, Hibberd JM, Kelly S, Covshoff S, Wong GKS, Hancock L, et al. 2014. Shared origins of a key enzyme during the evolution of C_4 and CAM metabolism. *J Exp Bot.* 65:3609–3621.

Christin PA, Boxall SF, Gregory R, Edwards EJ, Hartwell J, Osborne CP. 2013. Parallel recruitment of multiple genes into C4 photosynthesis. *Genome Biol Evol.* 5:2174–2187.

Christin PA, Osborne CP. 2013. The recurrent assembly of C_4 photosynthesis, an evolutionary tale. *Photosyn Res.* 117:163–175.

Christin PA, Osborne CP. 2014. The evolutionary ecology of C_4 plants. *New Phytol.* 204:765–781.

Christin PA, Osborne CP, Chatelet DS, Columbus JT, Besnard G, Hodkinson TR, Garrison LM, Vorontsova MS, Edwards EJ. 2013. Anatomical enablers and the evolution of C_4 photosynthesis in grasses. *Proc Natl Acad Sci U S A.* 110:1381–1386.

Christin PA, Osborne CP, Sage RF, Arakaki M, Edwards EJ. 2011. C_4 eudicots are not younger than C_4 monocots. *J Exp Bot.* 62: 3171–3181.

Dunn CW, Howison M, Zapata F. 2013. Agalma: an automated phylogenomics workflow. *BMC Bioinformatics* 14:330.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32:1792–1797.

Edwards EJ, Ogburn RM. 2012. Angiosperm responses to a low-CO_2 world: CAM and C_4 photosynthesis as parallel evolutionary trajectories. *Int J Plant Sci.* 173:724–733.

Edwards EJ, Smith SA. 2010. Phylogenetic analyses reveal the shady history of C_4 grasses. *Proc Natl Acad Sci U S A.* 107:2532–2537.

Fischer I, Dainat J, Ranwez V, Glémin S, Dufayard JF, Chantret N. 2014. Impact of recurrent gene duplication on adaptation of plant genomes. *BMC Plant Biol.* 14:151.

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

Fong SS, Joyce AR, Palsson BØ. 2005. Parallel adaptive evolution cultures of *Escherichia coli* lead to convergent growth phenotypes with different gene expression states. *Genome Res.* 15:1365–1372.

Gowik U, Brautigam A, Weber KL, Weber APM, Westhoff P. 2011. Evolution of C4 photosynthesis in the genus *Flaveria*: how many genes and which genes does it take to make C4? *Plant Cell* 23: 2087–2105.

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

Griffiths H, Weller G, Toy LFM, Dennis RJ. 2013. You're so vein: bundle sheath physiology, phylogeny and evolution in C_3 and C_4 plants. *Plant Cell Environ.* 36:249–261.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol.* 52: 696–704.

Harms MJ, Thornton JW. 2014. Historical contingency and its biophysical basis in glucocorticoid receptor evolution. *Nature* 512:203–207.

Hatch MD. 1987. C_4 photosynthesis: a unique blend of modified biochemistry, anatomy and ultrastructure. *Biochim Biophys Acta.* 895: 81–106.

Heckmann D, Schulze S, Denton A, Gowik U, Westhoff P, Weber AP, Lercher MJ. 2013. Predicting C_4 photosynthesis evolution: modular, individually adaptive steps on a Mount Fuji fitness landscape. *Cell* 7: 1579–1588.

Hibberd JM, Covshoff S. 2010. The regulation of gene expression required for C_4 photosynthesis. *Annu Rev Plant Biol.* 61:181–207.

Hibberd JM, Quick WP. 2002. Characteristic of C_4 photosynthesis in stems and petioles of C_3 flowering plants. *Nature* 415:451–454.

John CR, Smith-Unna RD, Woodfield H, Hibberd JM. 2014. Evolutionary convergence of cell specific gene expression in independent lineages of C_4 grasses. *Plant Physiol.* 165:62–75.

Kadereit G, Ackerly D, Pirie MD. 2012. A broader model for C_4 photosynthesis evolution in plants inferred from the goosefoot family (Chenopodiaceae s.s.). *Proc Biol Sci.* 279:3304–3311.

Kajala K, Brown NJ, Williams BP, Borrill P, Taylor LE, Hibberd JM. 2012. Multiple *Arabidopsis* genes primed for recruitment into C_4 photosynthesis. *Plant J.* 69:47–56.

Kraybill AA, Martin CE. 1996. Crassulacean acid metabolism in three species of the C_4 genus *Portulaca*. *Int J Plant Sci.* 157:103–109.

Külahoglu C, Denton AK, Sommer M, Maß J, Schliesky S, Wrobel TJ, Berckmans B, Gongora-Castillo E, Buell CR, Simon R, et al. 2014. Comparative transcriptome atlases reveal altered gene expression modules between two Cleomaceae C_3 and C_4 plant species. *Plant Cell* 26:3243–3260.

Langmead B, Salzberg S. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods.* 9:357–359.

Lara MV, Disante KB, Podesta FE, Andreo CS, Drincovich MF. 2003. Induction of a Crassulacean acid like metabolism in the C_4 succulent plant, *Portulaca oleracea* L.: physiological and morphological changes are accompanied by specific modifications in phosphoenolpyruvate carboxylase. *Photosyn Res.* 77:241–254.

Mallmann J, Heckmann D, Bräutigam A, Lercher MJ, Weber AP, Westhoff P, Gowik U. 2014. The role of photorespiration during the evolution of C_4 photosynthesis in the genus *Flaveria*. *eLife* 3: e02478.

Mallona I, Egea-Cortines M, Weiss J. 2011. Conserved and divergent expression rhythms of CAM related and core clock genes in the cactus *Opuntia ficus-indica*. *Plant Physiol.* 156:1978–1989.

Marazzi B, Ané C, Simon MF, Delgado-Salinas A, Luckow M, Sanderson MJ. 2012. Locating evolutionary precursors on a phylogenetic tree. *Evolution* 66:3918–3930.

Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. 2008. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.* 18:1509–1517.

McGonigle B, Nelson T. 1995. C_4 isoform of NADP-malate dehydrogrenase: cDNA cloning and expression in leaves of C_4, C_3, and C_3-C_4 intermediate species of *Flaveria*. *Plant Physiol.* 108: 1119–1126.

Monson RK. 2003. Gene duplication, neofunctionalization, and the evolution of C_4 photosynthesis. *Int J Plant Sci.* 164:S43–S54.

Muhaidat R, McKown AD. 2013. Significant involvement of PEP-CK in carbon assimilation of C_4 eudicots. *Ann Bot.* 111:577–589.

Muhaidat R, Sage RF, Dengler NG. 2007. Diversity of Kranz anatomy and biochemistry in C_4 eudicots. *Am J Bot.* 94:362–381.

Osborne CP, Freckleton RP. 2009. Ecological selection pressures for C_4 photosynthesis in the grasses. *Proc R Soc B.* 276:1753–1760.

Osmond CB. 1978. Crassulacean acid metabolism. A curiosity in context. *Annu Rev Plant Physiol.* 29:379–414.

Rensing SA. 2014. Gene duplication as a driver of plant morphogenetic evolution. *Curr Opin Plant Biol.* 17:43–48.

Rosche E, Westhoff P. 1990. Primary structure of pyruvate, orthophosphate dikinase in the dicotyledonous C_4 plant *Flaveria trinervia*. *FEBS Lett.* 273:116–121.

Sage RF. 2001. Environmental and evolutionary preconditions for the origin and diversification of the C_4 photosynthetic syndrome. *Plant Biol.* 3:202–213.

Sage RF. 2004. The evolution of C_4 photosynthesis. *New Phytol.* 161: 341–370.

Sage RF, Christin P-A, Edwards EJ. 2011. The C_4 plant lineages of planet Earth. *J Exp Bot.* 62:3155–3169.

Sage RF, Sage TL, Kocacinar F. 2012. Photorespiration and the evolution of C_4 photosynthesis. *Annu Rev Plant Biol.* 63:19–47.

Sawyer SA. 1999. GENCONV: a computer package for the statistical detection of gene conversion. Distributed by the author, Department of Mathematics, Washington University in St Louis. Available from http://www.math.wustl.edu/~sawyer.

Siebert S, Robinson MD, Tintori SC, Goetz F, Helm RR, Smith SA, Shaner N, Haddock SHD, Dunn CW. 2011. Differential gene expression in the siphonophore *Nanomia bijuga* (Cnidaria) assessed with multiple next-generation sequencing workflows. *PLoS One* 6: e22953.

Smith SA, Beaulieu JM, Donoghue MJ. 2010. An uncorrelated relaxed-clock analysis suggests an earlier origin for flowering plants. *Proc Natl Acad Sci U S A.* 107:5897–5902.

Soltis DE, Smith SA, Cellinese N, Wurdack KJ, Tank DC, Brockington SF, Refulio-Rodriguez NF, Walker JB, Moore MJ, Carlsward BS, et al. 2011. Angiosperm phylogeny: 17 genes, 640 taxa. *Am J Bot.* 98: 704–730.

Svensson P, Bläsing OE, Westhoff P. 2003. Evolution of C_4 phosphoenolpyruvate carboxylase. *Arch Biochem Biophys.* 414:180–188.

Tanz SK, Tetu SG, Vella NG, Ludwig M. 2009. Loss of the transit peptide and an increase in gene expression of an ancestral chloroplastic carbonic anhydrase were instrumental in the evolution of the cytosolic C_4 carbonic anhydrase in *Flaveria*. *Plant Physiol.* 150: 1515–1529.

Tausta SL, Miller Coyle H, Rothermel B, Stifel V, Nelson T. 2002. Maize C_4 and non-C_4 NADP-dependent malic enzymes are encoded by distinct genes derived from a plastid-localized ancestor. *Plant Mol Biol.* 50:635–652.

Thompson JD, Higgins DJ, Gibson TJ. 1994. ClustalW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and matrix choice. *Nucleic Acids Res.* 22:4673–4680.

Tronconi MA, Fahnenstich H, Weehler MCG, Andreo CS, Flügge UI, Drincovich MF, Maurino VG. 2008. *Arabidopsis* NAD-malic enzyme functions as a homodimer and heterodimer and has a major impact on nocturnal metabolism. *Plant Physiol.* 146: 1540–1552.

Wang X, Gowik U, Tang H, Bowers JE, Westhoff P, Paterson AH. 2009. Comparative genomic analysis of C_4 photosynthetic pathway evolution in grasses. *Genome Biol.* 10:R68.

Weinreich DM, Delaney NF, DePristo MA, Hartl DL. 2006. Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science* 312:111–114.

Williams BP, Aubry S, Hibberd JM. 2012. Molecular evolution of genes recruited into C_4 photosynthesis. *Trends Plant Sci.* 17:213–220.

Williams BP, Johnston IG, Covshoff S, Hibberd JM. 2013. Phenotypic landscape inference reveals multiple evolutionary paths to C_4 photosynthesis. *eLife* 2:e00961.

Winter K, Holtum JAM. 2014. Facultative crassulacean acid metabolism (CAM) plants: powerful tools for unravelling the functional elements of CAM photosynthesis. *J Exp Bot.* 65:3425–3441.

Zhang J. 2003. Evolution by gene duplication: an update. *Trends Ecol Evol.* 18:292–298.