RESEARCH PAPER

Shared origins of a key enzyme during the evolution of C₄ and CAM metabolism

Pascal-Antoine Christin¹,²*, Monica Arakaki²,³, Colin P. Osborne¹, Andrea Bräutigam⁴, Rowan F. Sage⁵, Julian M. Hibberd⁶, Steven Kelly⁷, Sarah Covshoff⁶, Gane Ka-Shu Wong⁸,⁹,¹⁰, Lillian Hancock² and Erika J. Edwards²

¹ Department of Animal and Plant Sciences, University of Sheffield, Western Bank, Sheffield S10 2TN, UK
² Department of Ecology and Evolutionary Biology, Brown University, 80 Waterman St., Providence, RI 02912, USA
³ Departamento de Botánica, Facultad de Ciencias Biológicas and Museo de Historia Natural – UNMSM, Av. Arenales 1256, Lima 11, Peru
⁴ Institute of Plant Biochemistry, Heinrich-Heine University, 40225 Duesseldor, Germany
⁵ Department of Ecology and Evolutionary Biology, University of Toronto, 25 Willcocks Street, Toronto, Ontario M5S 3B2, Canada
⁶ Department of Plant Sciences, University of Cambridge, Cambridge CB2 3EA, UK
⁷ Department of Plant Sciences, University of Oxford, Oxford OX1 3RB, UK
⁸ Department of Biological Sciences, University of Alberta, Edmonton, Alberta T6G 2E9, Canada
⁹ Department of Medicine, University of Alberta, Edmonton, Alberta T6G 2E1, Canada
¹⁰ BGI-Shenzhen, Beishan Industrial Zone, Yantian District, Shenzhen 518083, China

* To whom correspondence should be addressed. E-mail: p.christin@sheffield.ac.uk

Received 29 November 2013; Revised 31 January 2014; Accepted 3 February 2014

Abstract

CAM and C₄ photosynthesis are two key plant adaptations that have evolved independently multiple times, and are especially prevalent in particular groups of plants, including the Caryophyllales. We investigate the origin of photosynthetic PEPC, a key enzyme of both the CAM and C₄ pathways. We combine phylogenetic analyses of genes encoding PEPC with analyses of RNA sequence data of Portulaca, the only plants known to perform both CAM and C₄ photosynthesis. Three distinct gene lineages encoding PEPC exist in eudicots (namely ppc-1E1, ppc-1E2 and ppc-2), one of which (ppc-1E1) was recurrently recruited for use in both CAM and C₄ photosynthesis within the Caryophyllales. The PEPC involved in the CAM and C₄ cycles of Portulaca are encoded by closely related yet distinct genes. The CAM-specific gene is similar to genes from related CAM taxa, suggesting that CAM has evolved before C₄ in these species. The similar origin of PEPC and other genes involved in the CAM and C₄ cycles highlights the shared early steps of evolutionary trajectories towards CAM and C₄, which probably diverged irreversibly only during the optimization of CAM and C₄ phenotypes.

Key words: C₄ photosynthesis, CAM photosynthesis, co-option, evolution, phosphoenolpyruvate carboxylase (PEPC), phylogenetics.

Introduction

During the evolutionary diversification of organisms, ecological selection pressures sometimes lead to the emergence of similar phenotypes in distantly related species. A good example of such convergent evolution is the recurrent emergence of CO₂-concentrating mechanisms (CCMs) as an adaptation to environmental CO₂ depletion (Raven et al., 2008). CCMs have arisen through the assembly of novel biochemical pathways, which increase the internal concentration of CO₂ around Rubisco before its fixation by the C₃ photosynthetic cycle (Christin and Osborne, 2013). In flowering plants, the most
frequent and successful CCMs are C₄ and CAM photosynthesis (Keeley and Rundel, 2003). The C₄ and CAM CCMs differ in the overall mechanism of atmospheric CO₂ concentration, but the biochemical cycles are similar (Osmond, 1978; Hatch, 1987). Plants of both types fix inorganic carbon by phosphoenolpyruvate carboxylase (PEPC). In C₄ plants, the resulting acid is typically modified and transported to another cell, where CO₂ is released to feed the C₃ cycle that is active within these cells (Hatch, 1987; Sage et al., 2012). In CAM plants, a similar CCM occurs in which the initial carboxylation by PEPC and subsequent decarboxylation and refixation by Rubisco are temporally rather than physically separated. At night, PEPC fixes inorganic carbon into organic compounds, which are stored as malate in the vacuole until daytime, when the cycle is completed and CO₂ is released to supply the C₃ cycle (Osmond, 1978; Borland and Taybi, 2004).

The CAM and C₄ pathways are complex traits, involving dozens of enzymes that fulfill different functions compared with the isoforms in the C₃ ancestors (Bräutigam et al., 2011; Gowik et al., 2011). For the constituent enzymes investigated so far, the new function requires alterations in the expression pattern and/or its catalytic properties (Finnegan et al., 1999; Tausta et al., 2002; Svensson et al., 2003; Gowik and Westhoff, 2011; Ludwig, 2011). For instance, the PEPC enzyme is ubiquitous in plants, where multiple isoforms are responsible for various non-photosynthetic functions (Lepiniec et al., 1994; Aubry et al., 2011; Gowik and Westhoff, 2011). The C₄-specific forms evolved from non-photosynthetic genes through adaptation of the catalytic properties to the new metabolic context (Dong et al., 1998; Gowik and Westhoff, 2011). In particular, the affinity of PEPC for its substrate PEP was decreased, and its sensitivity to feedback inhibition by malate was reduced (Bauwe and Chollet, 1986; Bläsig et al., 2000; Svensson et al., 2003; Gowik et al., 2006; Lara et al., 2006). In C₄ grasses and sedges, this was achieved through numerous adaptive amino acid changes (Christin et al., 2007; Besnard et al., 2009).

Both the CAM and C₄ pathways also require a specialized leaf anatomy (Hattersley, 1984; Nelson et al., 2005). Despite this apparent complexity, the C₄ trait evolved a minimum of 62 times independently in flowering plants (Sage et al., 2011). In particular, the affinity of PEPC for its substrate PEP was decreased, and its sensitivity to feedback inhibition by malate was reduced (Bauwe and Chollet, 1986; Bläsig et al., 2000; Svensson et al., 2003; Gowik et al., 2006; Lara et al., 2006). In C₄ grasses and sedges, this was achieved through numerous adaptive amino acid changes (Christin et al., 2007; Besnard et al., 2009). The CAM and C₄ origins might suggest that this clade is especially prone to transitions between photosynthetic types. However, the history of photosynthetic transitions within the clade is still unclear. In particular, it is unknown if CAM and C₄ origins represent completely independent evolutionary phenomena, or are distinct end-points to a partially shared evolutionary trajectory (Edwards and Ogburn, 2012). To increase our understanding of CCM origins in Caryophyllales, we studied the evolution of genes encoding PEPC, a key enzyme common to all CAM and C₄ cycles. In addition, we investigated in detail the transcriptome of Portulaca individuals operating either the C₄ or CAM cycles. Our comparative analyses shed new light on the shared history of genes involved in CAM and C₄ photosynthesis.

**Material and methods**

**Diversity of genes encoding PEPC in plants**

To reconstruct the history of the multigene family encoding PEPC, complete cDNA sequences available in the GenBank database were first retrieved. These were used as a query of BLAST searches against completely sequenced nuclear genomes available on Phytozome (Goodstein et al., 2012). This initial dataset was increased by screening genomic DNA from representatives of diverse Caryophyllales lineages and photosynthetic types through PCR to isolate genes encoding PEPC (Supplementary Tables S1 and S2, available at JXB online). These DNAs were first screened with primers ppc-1204-For and ppc-2890-Rev previously used to isolate PEPC genes from Molluginaceae, a family from Caryophyllales (Christin et al., 2011). These primers amplify a fragment encompassing exons 8 to 10, which represents about half of the whole coding sequence and is known to include...
major determinants of the C₄-specific properties of PEPC (Bläsing et al., 2000; Engelmann et al., 2002). PCR, cloning and sequencing were performed as previously described (Christin et al., 2011). Preliminary analyses indicated that numerous PEPC genes were present in some Caryophyllales genomes. To increase the likelihood of sampling specific copies, additional primers were designed with the purpose of increasing the PCR specificity for certain gene lineages (Supplementary Table S3, available at JXB online). These additional PCRs were conducted as for the primers ppc-1204-For and ppc-2890-Rev. The PCR products were purified and directly sequenced with one of the primers used for the PCR. Chromatograms were visually inspected and PCR products were cloned only if multiple genes were detected by the presence of overlapping peaks.

Exons were identified by homology to annotated sequences and following the GT-AG rule. All coding sequences, retrieved from public databases or isolated through PCR, were translated into protein sequences and aligned with ClustalW (Thompson et al., 1994). The alignment was visually inspected, manually refined, and replaced with the corresponding nucleotide sequences, which were later used for analyses. A preliminary phylogenetic tree identified two distantly related groups of PEPC encoding genes that diverged before the evolution of land plants (ppc-1 and ppc-2; Fig. 1), each represented by ferns, gymnosperms, and flowering plants. Despite clear homology, ppc-1 and ppc-2 were highly divergent, leading to ambiguities in the alignment. Each group was consequently analysed separately.

Transcriptome analysis of Portulaca oleracea

The species Portulaca oleracea was selected for transcriptome analyses to identify CAM- and C₄-specific genes, with a special focus on PEPC. Seeds originating from Syria were provided by the USDA-ARS (GRIN accession: Griff 14515). Plants were grown in 3-inch pots of equal-parts gravel/calcined clay perlite mix, and in a Conviron E7/2 plant growth chamber (Conviron Ltd., Winnipeg, Manitoba, Canada) with 14 hours of daylight. The growth chamber was illuminated with twelve 32-W fluorescent lamps and four 60-W incandescent lamps. Temperature was kept at 22°C from 3 h before dark until after 3 h of light, and was increased to 28°C for the middle portion of the light period. The position of pots within the growth chamber was randomized daily.

On the first day of the experiment (5 March 2012), all plants were bottom-watered and seedlings were split into two groups. One group was bottom-watered every 2–4 days, while the other group were bottom-watered and seedlings were split into two groups. One individual of each group was sampled first at the end of the dark period (see below). One individual of each group was sampled first at the end of the dark period (1 h before light) and then at the end of the light cycle (2 h before dark). The sampling order was inverted for the middle portion of the light period. The assignment of contigs to gene lineages was repeated with a reference dataset composed of exons 8–10 of genes encoding PEPC. Caryophyllales sequences isolated by PCR or retrieved from Genbank were used if complete for the studied fragment and assigned to the gene lineage ppc-1E1 based on phylogenetic analyses (see Results). In order to increase the size of this dataset, the corresponding segment of matching contigs from Portulaca transcripts that represented at least half of the studied fragment were manually aligned with the dataset. Three contigs that covered the whole studied segment and which were clearly different from each other, as well as from Portulaca genes isolated by PCR, were added to the reference dataset. All matching Portulaca contigs were then successively placed in a phylogeny with this reference dataset and the total rpm values of each ppc-1E1 gene lineage was computed as described above.

Titratable acidity

The titratable acidity was measured at the end of the dark phase and at the end of the light phase. An accumulation of acids during the night followed by consumption during the day is indicative of a CAM cycle (Silvera et al., 2005). Three leaves were analysed for each of the eight combinations of individuals/sampling time. Frozen leaves were weighed and immediately ground briefly in mortar and pestle and then boiled in 50 ml 20% ethanol. After cooling, pH was slowly brought to 7 by adding 0.1 N NaOH in 5 or 10 μl increments. The titratable acidity was calculated as the amount of H⁺ equivalents required to neutralize the leaf extracts (Silvera et al., 2005). Despite variability among individuals and leaves, the plants watered occasionally showed a clear increase of titratable acidity at the end of the dark period (Fig. 2). These results confirm that less frequent watering triggered a CAM physiology in Portulaca oleracea.

Illumina data assembly and phylogenetic annotation

The reads corresponding to each of the eight different samples were assembled individually using the software Trinity (Grabherr et al., 2011) as implemented in the Agalma pipeline (Dunn et al., 2013). Reads from each Illumina run were mapped against the corresponding individual assembly using the software Bowtie 2 (Langmead and Salzberg, 2012) and its mixed model, which allows unpaired alignments when paired alignments fail. Only one of the best alignments chosen randomly was reported per read. The number of reads aligned to each contig was used to compute reads per million of alignable reads (rpm). Sequencing and mapping statistics are given in Supplementary Table S5 (available at JXB online).

The relative transcript abundance of each gene within each of the C₄-related multigene families was estimated by assigning contigs to gene lineages based on phylogenetic analyses (Christin et al., 2013a). For genes encoding C₄-related enzymes (Supplementary Table S6, available at JXB online), sequences available in Genbank for Caryophyllales and other plant lineages were extracted based on their annotation. A BLAST search was used to extract homologous loci for predicted cDNA from several completely sequenced genomes (Arabidopsis thaliana, Brachypodium distachyon, Carica papaya, Glycine max, Oryza sativa, Populus trichocarpa, Selaginella moellendorffii, Sorghum bicolor, Selaginella moellendorffii and Vitis vinifera). For each C₄-related enzyme, the dataset assembled from sequences retrieved from Genbank and complete genomes was used as a query of a BLAST search against each of the assembled transcriptomes. The longest matching region for each contig was extracted from the BLAST results when larger than 50 bp. Each of these was successively aligned to the reference dataset using Muscle (Edgar, 2004). The resulting alignment was used to infer one phylogenetic tree per contig under maximum likelihood as implemented in PhyML (Guindon and Gascuel, 2003) using a GTR+G model. The resulting phylogenetic trees were inspected and each contig was assigned to a gene lineage if clearly nested within. The rpm values of all contigs assigned to a given gene lineage were summed to obtain an estimate of the transcript abundance of the gene lineage.

The assignment of contigs to gene lineages was repeated with a reference dataset composed of exons 8–10 of genes encoding PEPC. Caryophyllales sequences isolated by PCR or retrieved from Genbank were used if complete for the studied fragment and assigned to the gene lineage ppc-1E1 based on phylogenetic analyses (see Results). In order to increase the size of this dataset, the corresponding segment of matching contigs from Portulaca transcripts that represented at least half of the studied fragment were manually aligned with the dataset. Three contigs that covered the whole studied segment and which were clearly different from each other, as well as from Portulaca genes isolated by PCR, were added to the reference dataset. All matching Portulaca contigs were then successively placed in a phylogeny with this reference dataset and the total rpm values of each ppc-1E1 gene lineage was computed as described above.
Fig. 1. Phylogenetic relationships among PEPC-encoding genes from land plants. The phylogenetic trees were obtained through Bayesian inference on each main gene lineage; ppc-1 and ppc-2. Taxonomic groups are compressed, with the size of triangles proportional to the number of sequences in the group. Gene lineages and main clades of flowering plants are delimited on the right. Clades containing some genes with a Ser780 are in orange. Details, including support values, are available in Supplementary Figs S1 and S2 (available at JXB online). The scale bar represents expected substitutions per sites. (This figure is available in colour at JXB online.)
Phylogenetic analyses and codon models

The two datasets composed of all ppc-1 and ppc-2 genes retrieved from available databases or generated in this study were used to infer phylogenetic trees with MrBayes 3.2.1 (Ronquist and Huelsenbeck, 2003). The general time reversible model of nucleotide substitution with a gamma-shape parameter and a proportion of invariants (GTR+G+I) was used. For ppc-2, two parallel analyses each composed of four chains were run for 2000000 generations, sampling a tree every 1000 generations after a burn-in period of 1000000 generations. Due to slow convergence of parallel runs the number of parallel chains was increased to sixteen for analyses of ppc-1. For this dataset, two different analyses were run for 10000000 generations, sampling a tree every 1000 generations after a burn-in period of 4000000 generations. Consensus trees were computed from all trees sampled after the burn-in period. Convergence of the analyses and adequacy of the burn-in period were determined using Tracer (Rambaut and Drummond, 2007).

In order to represent the rate of amino acid changes, the topologies inferred from the whole ppc-1 and ppc-2 datasets were used to infer branch lengths based on amino acid sequences while keeping the topology constant. These analyses were performed in PhyML, under a JTT+G substitution model. Statistical tests of adaptive evolution were also conducted on the group of ppc-1E1 sequences from Caryophyllales using codon models implemented in codeml of the PAML package (Yang, 2007). These models use the ratio of non-synonymous mutation rate per synonymous mutation rate (ω) as a proxy of selective pressures. An ω smaller than 1 indicates purifying selection, a value of 1 indicates relaxed selection, and a value greater than 1 indicates adaptive evolution. Different models allow ω to vary among sites of the protein or among both sites and branches of the phylogeny (Yang and Nielsen, 2002). The site model without adaptive evolution (M1a) was compared with the site model assuming adaptive evolution on some sites but on all branches (M2a), as well as to several branch-site models assuming adaptive evolution only on some sites and on some branches (referred to as foreground branches; model A). In these branch-site models, foreground branches have to be defined a priori. Different sets of foreground branches were successively selected, and the likelihoods were compared using Akaike information criteria (AIC). In this first model (a), foreground branches were defined as each branch leading to a group of putative C₄ forms, to the group of sequences belonging to the CAM genera Mesembryanthemum and Drosanthemum, or to the group of sequences belonging to Portulaca and present at high transcript abundance in the night samples (putative CAM form; see Results). In the other models, entire gene lineages present in Portulacinae were successively added to the foreground branches: (b) branches a + ppc-1E1c, (c) branches b + ppc-1E1d + ppc-1E1e, (d) branches c + ppc-1E1b, and (e) branches d + ppc-1E1a.

Since the evolutionary history of C₄ photosynthesis in the genus Portulaca is debated (Ocampo et al., 2013), the group of Portulaca sequences that contains putative C₄ forms (ppc-1E1a' and ppc-1E1a) was analysed in more detail. Parallel adaptive genetic changes can bias the phylogenetic reconstruction when they occur in closely related taxa, and considering only third positions of codons can help recover the true phylogenetic relationships (Christin et al., 2007, 2012). To avoid a potential bias due to adaptive evolution, a phylogenetic tree was inferred as described above but considering only third positions of codons from Portulaca ppc-1E1a and ppc-1E1a’ sequences. The inferred topology was used to test alternative scenarios of adaptive evolution using the codon models described above. In addition, to the site models M1a and M2a, different branch-site models (A) were compared. Foreground branches were set following three evolutionary scenarios: adaptive evolution at the base of Portulaca (on the branch leading to ppc-1E1a'; single C₄ optimization), at the base of each C₄ clade of Portulaca ppc-1E1a’ (multiple C₄ optimizations); and finally at the base of each C₄ and C₃–C₄ clade of Portulaca ppc-1E1a’ (multiple C₄ and C₃–C₄ optimizations).

Results

PEPC multigene family

The phylogenetic tree inferred using all genes encoding PEPC retrieved from GenBank, the 1KP project, or generated in this study showed that a gene duplication occurred before the emergence of land plants and produced two groups of distantly related genes present in most plant genomes (ppc-1 and ppc-2; Fig. 1; Gowik and Westhoff, 2011). One of these groups (ppc-1) contains all the C₄-specific PEPC genes documented so far (Rao et al., 2002; Svensson et al., 2003; Gowik et al., 2006; Christin et al., 2007, 2011; Besnard et al., 2009; Gowik and Westhoff, 2011). The phylogenetic relationships in each group are compatible with the species relationships predicted from other markers (Supplementary Figs S1 and S2, available at JXB online; Soltis et al., 2011). The gene
ppc-2 is present as a single copy in all the species considered (Fig. S2). No gene duplication of ppc-1 is detectable before the split of eudicots and monocots, but several gene duplications led to the six gene lineages present in grass genomes as shown previously (namely ppc-aL1a, ppc-aL1b, ppc-aL2, ppc-aR, ppc-B1, and ppc-B2; Fig. S1; Christin and Besnard 2009). In addition, a gene duplication occurred soon after the early diversification of eudicots, leading to two gene lineages present in most eudicots and named ppc-1E1 and ppc-1E2 (Fig. 1 and Supplementary Fig. S1, available at JXB online; these correspond to ppc-1 and ppc-2 in Christin et al., 2011).

**Diversification of ppc-1E1 in Caryophyllales**

Genes belonging to the three ppc gene lineages present in eudicots were isolated from core Caryophyllales through PCR screening of genomic DNA (Supplementary Tables S1 and S2, available at JXB online). The phylogenetic trees of Caryophyllales ppc-2 and ppc-1E2 were consistent with published phylogenies based on chloroplast and nuclear markers (Brockington et al., 2009; Arakaki et al., 2011; Kadereit et al., 2012) and no ancient gene duplication was detectable (Supplementary Figs S1 and S2, available at JXB online). On the other hand, ppc-1E1 is present in a high number of copies in members of the Portulacineae (namely ppc1E1a to ppc1E1e in Fig. 3 and Supplementary Fig. S1, available at JXB online), indicating that this gene lineage was repeatedly duplicated during the early diversification of this clade. The species relationships deduced from each of these gene lineages are consistent with those deduced from chloroplast markers (Arakaki et al., 2011). The ppc1E1a gene lineage includes sequences isolated previously from cDNA of two members of Portulacineae, *Pereskia aculatea* and *Selenicereus vitii* (Gowik et al., 2001).

The ppc-1E1 lineage contains genes encoding putative C4-specific forms in *Alternanthera* (Gowik et al., 2006), *Bienertia* and *Suaeda* (Lara et al., 2006), as well as genes encoding putative CAM-specific forms in *Mesembryanthemum* (Rickers et al., 1989). Most C4-specific PEPC genes studied previously encode a serine residue at the position homologous to position 780 of the maize PEPC protein (numbering based on *Zea mays* sequence CAA33317; hereafter referred to as Ser780), although this Ser780 is not necessary for the C4 function (Rao et al., 2008). In *Flaveria*, Ser780 has been shown to be a major determinant of the C4 properties of PEPC (Bläsing et al., 2000; Engelmann et al., 2002; Svensson et al., 2003). The homologous position is occupied by a conserved alanine in all characterized PEPC genes not involved in C4 photosynthesis (Svensson et al., 2003; Christin et al., 2007, 2011; Besnard et al., 2009). In the present dataset, a Ser780 is encoded by ppc-1E1 genes from the C4 taxa in Amaranthaceae, Aizoaceae, and Molluginaceae, while it is not encoded by homologous genes of C3 taxa of the same families and from other gene lineages (Fig. 3 and Supplementary Figs S1 and S2, available at JXB online). This supports earlier suggestions that ppc-1E1 encodes the C4-specific PEPC in numerous C4 Caryophyllales (Gowik et al., 2006; Christin et al., 2011; Gowik and Westhoff, 2011). However, the Ser780 is not encoded by ppc-1E1 genes from some C3 taxa in Nyctaginaceae and Aizoaceae.

Most members of Portulacineae have the ability to perform some degree of CAM (Guralnick et al., 2008; Nyffeler et al., 2008). The ppc-1Elc genes of Portulacineae encode a Ser780 in most species (Fig. 3). Sequences isolated by PCR from cDNA extracted from the stems of *Nopalea* and *Echinocereus* at night belong to this gene lineage and encode a Ser780, which indicates that proteins encoded by ppc-1Elc might be involved in the CAM pathway of cacti. This is further supported by the isolation of ppc-1Elc sequences from cDNA of another cactus (*Hylocereus undatus*; NCBI accession JF96382). However, the Ser780 residue is not encoded by ppc-1Elc genes in the sampled Montiaceae and in several of the sampled Didiereaceae (Supplementary Fig. S1, available at JXB online). In addition, the ppc-1Elc from *Ceraria* (Didiereaceae; Supplementary Fig. S1, available at JXB online) also encodes a Ser780. All other Portulacineae ppc-1E1 gene copies and all ppc-2 and ppc-1E2 (with the exception of some sequences from *Flaveria* known to be involved in C4 photosynthesis; Svensson et al., 2003) encode the Ala780 typical of non-C4 genes.

**Transcriptome from Portulaca oleracea**

Differences in estimates of transcript abundance of C4-related genes between individuals grown under the same watering regime were small (Supplementary Table S6, available at JXB online). In the well-watered samples, some of the gene lineages encoding the enzymes of the C4 pathway (β-CA, PEPC, NADP-MDH, NADP-MDH, ALA-AT, ASP-AT, PPDK, NAD-ME and NADP-ME) were present at high transcript abundance during the day (Supplementary Table S6, available at JXB online). With the exception of NADP-ME, these correspond to the enzymes postulated to play a role in the C4 pathway of *P. oleracea* (Lara et al., 2004). After 2h in the dark, the transcript abundance of these enzymes remained substantial, although in most cases it was strongly reduced compared with the day sample (Supplementary Table S6, available at JXB online). The only exception is the gene for ALA-AT, which was present at higher transcript abundance in the dark than in the light. In addition, one of the individuals showed a slight increase in the transcript abundance of one of the genes for NADP-MDH at night while its abundance was extremely low in the other individual (Supplementary Table S6, available at JXB online).

In the samples watered less frequently, genes encoding the same enzymes were present at high transcript abundance during the day, although the levels were generally lower than in the well-watered sample (Supplementary Table S6, available at JXB online). At night, genes encoding all C4-related enzymes were present at lower abundance, with the exception of genes for PEPC and NADP-MDH, for which the abundance of one gene lineage increased to reach levels equivalent to, or even higher than, those observed for the well-watered samples during the day (Table 1 and Supplementary Table S6, available at JXB online). This is consistent with the nocturnal part of the purported CAM cycle of *P. oleracea*, which is based on these two enzymes (Lara et al., 2004). However, the nocturnal part was assumed to also involve carbonic anhydrase, but the transcript abundance of this enzyme is strongly reduced at night although it remains high (Supplementary Table S6, available at JXB online).
Fig. 3. Evolution of ppc-1E1 in Caryophyllales. The topology was inferred on nucleotide sequences, but branch lengths were estimated based on amino acid sequences. The branch lengths inferred on nucleotide sequences, together with all species names and support values, are available in Fig. S1. Groups of genes encoding a Ser780 are highlighted by red branches. Branches where some sites underwent an excess of non-synonymous mutations according to the best model are thicker. Putative C₄ forms are delimited in green and putative CAM forms in blue. These were identified based either on transcript abundances in specific conditions, on the literature, or on an excess of amino acid changes in C₄/CAM species. Genes of C₄ or CAM taxa that represent putative non-photosynthetic duplicates are delimited in grey, those of C₃ taxa in white, and those of C₃–C₄ taxa in yellow. Families outside Portulacineae and gene lineages for Portulacineae are indicated on the left: N, Nyctaginaceae; Mollug, Molluginaceae; Amar, Amaranthaceae; Aiz, Aizoaceae. Subclades of interest are indicated on the right: P, Portulaca; C, cacti. The full phylogenetic tree is available in Fig. S3. The scale bar represents expected substitutions per site. (This figure is available in colour at JXB online.)
The incorporation of contigs from the *P. oleracea* samples into the densely sampled Caryophyllales dataset allowed us to estimate the transcript abundance of each *ppc-1E1* gene lineage (Table 1). In addition to *ppc-1E2* and *ppc-2*, four distinct *ppc-1E1* genes were isolated from the transcriptomes of *P. oleracea*, only one of which was also isolated from genomic DNA (*ppc-1E1b*). One of these four genes was clearly nested within *ppc-1E1c* and two in *ppc-1E1a* (Supplementary Fig. S1, available at *JXB* online). The phylogenetic relationships suggest a recent duplication of *ppc-1E1a* in *Portulaca* and one of the duplicates was named *ppc-1E1a'*. The genes *ppc-1*, *ppc-2*, *ppc-1E1a*, and *ppc-1E1b* were present at similarly low abundances in all samples (Table 1). By contrast, *ppc-1E1a'* was present at very high transcript abundances during the day in the well-watered samples (Table 1). This pattern is consistent with a function in the C₄ pathway, which is moreover supported by the Ser780 encoded by the gene. The abundance of *ppc-1E1a'* during the day decreased in one of the individuals that were watered less frequently and its abundance at night decreased in both individuals (Table 1). The gene *ppc-1E1c* was present at extremely low transcript abundances in the well-watered samples during the day. However, its abundance increased at night, and the nocturnal abundance was considerably higher in infrequently watered than in well-watered plants (Table 1). High nocturnal transcript abundance triggered by reduced water availability supports an involvement of the encoded enzyme in the CAM pathway of *P. oleracea*. This gene also encodes a Ser780.

**Distribution of *ppc-1E1* genes in other Portulaca species and evidence of adaptive evolution**

Using the 1KP transcriptome data, sequences corresponding to the genes *ppc-1E1a* and *ppc-1E1c* were retrieved from the Oleracea clade, while sequences corresponding to *ppc-1E1b* were retrieved from both Oleracea and Pilosa clades (Supplementary Fig. S1, available at *JXB* online). The RNA for this project was isolated during the day and putative CAM-specific genes would likely be missed. The putative C₄ gene *ppc-1E1a'* was retrieved from the four *Portulaca* clades. One sequence attributed to *P. suffruticosa* was nested within *ppc-1E1a'*. The Pilosa clade (Fig. 4 and Supplementary Fig. S1, available at *JXB* online), which might indicate a biologically relevant phenomenon (e.g. hybridization) or a methodological problem (e.g. cross-contamination). Since these sequences were retrieved from leaf RNA isolated during the day, the presence of transcripts corresponding to *ppc-1E1a'* is compatible with the hypothesis that this gene is involved in C₄ photosynthesis of these different species.

The branch lengths estimated from amino acid sequences strongly vary among clades (Fig. 3 and Supplementary Fig. S3, available at *JXB* online). Since the variation is more pronounced than with nucleotide sequences (Supplementary Figs S1 and S2, available at *JXB* online), it indicates an excess of non-synonymous mutations. In Molluginaceae and Amaranthaceae, clear increases in the rate of amino acid substitutions occurred on branches leading to genes encoding putative C₄-specific enzymes. In Aizoaceae, a similar acceleration is visible on the branch leading to the putative C₄ gene of *Trianthema*, but also to the genes of the CAM plant *Mesembryanthemum crystallinum*. In addition, the branch leading to the C₄ Nyctaginaceae *Boerhavia* underwent many amino acid changes (Fig. 3), which is suggestive of functional divergence, potentially as an adaptation to the C₄ context. These *ppc-1E1* genes likely encode proteins involved in CCMs, but do not encode the Ser780. This indicates that changes happened in different parts of the coding sequences. The monophyletic group composed of Portulacineae *ppc-1E1c*, *ppc-1E1d*, and *ppc-1E1e* is also characterized by increased rates of amino acid substitutions, with further increases in some branches, such as those leading to *ppc-1E1c* from *Portulaca*, but also the cacti (Fig. 3 and Supplementary Fig. S3, available at *JXB* online). Within Portulacineae *ppc-1E1a'* long branches lead to the putative C₄-specific genes of *Portulaca*. However, most of the amino acid substitutions occurred after the divergence of the four clades (*sensu* Ocampo and Columbus, 2012) and comparatively few happened on the branch leading to the C₃–C₄ *P. cryptopetala* (Fig. 3).

The action of adaptive evolution on some branches of the phylogeny is supported by codon models. While the model assuming positive selection on some sites but all branches of Caryophyllales a *ppc-1E1* was not better than the null model, assuming increased rates of non-synonymous changes on
some sites but some branches only led to a very significant increase in likelihood (Table 2). While all the tested sets of foreground branches led to a significant increase of likelihood, the model assuming increased rates of non-synonymous mutations in the whole of Portulacineae ppc-1E1c, ppc-1E1d, and ppc-1E1e gene lineages in addition to C4 and CAM clades outside of Portulacineae produced the best AIC (Table 2). In this model, 16.8% of sites were estimated to undergo more non-synonymous mutations in the selected foreground branches, although the optimized dN/dS ratio was not different from 1. The model assuming adaptive evolution on branches leading to both C4 and C3–C4 clades was also better than the null model, but not as good as the model without adaptive evolution on the C3–C4 branch (difference of AIC >18).

**Discussion**

**Increased rates of amino acid changes in both C4 and CAM origins**

The evolution of genes encoding PEPC in the Caryophyllales is characterized by increased rates of amino acid substitutions and the recurrence of several amino acid changes previously detected in C4 monocots (e.g. E572Q, H665N, and A780S; Christin et al., 2007; Besnard et al., 2009). These increased rates of amino acid change are not limited to C4 taxa, but are also observed in CAM lineages, including Aizoaceae and Portulacineae species (Fig. 3), and the excess of non-synonymous mutations on these branches was confirmed with codon models (Table 2). C4 and CAM-specific PEPC differ in the timing of their activity, but the catalytic challenges they face in the two cycles are similar, because in both cases the concentrations of both substrates and products are greatly increased. The evolution of both C4- and CAM-specific PEPC consequently required adaptive mutations, some of which are shared among multiple origins, while several are probably specific to one or a few clades and might depend on the other amino acid mutations undergone by the coding sequence of the gene before its co-option.

In Portulacineae, the ppc-1E1 gene lineage is present in five copies, which appeared through several rounds of gene evolution.
duplications (Fig. 3 and Supplementary Fig. S1, available at JXB online). The three most recent copies, namely ppc-1E1c, ppc-1E1d, and ppc-1E1e, are all characterized by increased rates of amino acid substitutions (Fig. 3 and Table 2). The Portulacineae encompass species with different degrees of CAM metabolism (Guralnick and Jackson, 2001; Nyffeler et al., 2008). The high number of ppc-1E1 copies could have promoted neofunctionalization of the genes by relaxing selective constraints, facilitating the diversification of photosynthetic types in this group. A gradual upregulation of the CCM over time would have triggered successive periods of adaptive genetic changes in response to modifications of the catalytic environment, explaining the high rates of amino acid substitutions sustained in the entire clade (Fig. 3). For instance, the accumulation of mutations on the branches leading to ppc-1E1c of CAM-constitutive cacti (Cactoideae and Opuntioideae; Fig. 3) could be linked to the evolution of a more efficient CAM pathway in these taxa. This contrasts with the evolution of C4-specific PEPC where adaptive changes are concentrated at the base of each C4 group (Fig. 3; Christin et al., 2007; Besnard et al., 2009), and might indicate that the optimization of PEPC for the CAM function is spread over a longer time period.

C4 origins in Portulaca within a CAM-like context

The putative gene encoding the CAM-specific PEPC in Portulaca belongs to the ppc-1E1c gene lineage (Fig. 3). This gene lineage is characterized by increased rates of amino acid substitutions in other CAM taxa, such as cacti. This suggests that members of this ppc-1E1c gene lineage may have been already involved in some type of CAM metabolism before the divergence of Portulaca and cacti. On the other hand, the putative C4-specific genes of Portulaca belong to the ppc-1E1a gene lineage, which is duplicated in these taxa. One of the duplicates was likely co-opted for C4 photosynthesis after the gene duplication. Other members of the ppc-1E1a gene lineage, including the second duplicate of Portulaca, underwent mutations that generated amino acid substitutions at the same rate as genes from C3 species (Fig. 3). Codon models confirmed that adaptive non-synonymous mutations did not occur on these genes, but were restricted to some members of the ppc-1E1a′ duplicate, which is specific to Portulaca. The evolution of C4-specific genes in Portulaca likely co-opted a non-CCM gene through numerous changes in the coding sequences. Therefore, the distribution of high rates of non-synonymous substitutions indicates that the evolution of C4-specific PEPC occurred after the divergence of Portulaca from other Portulacineae while the CAM-specific properties of the gene used by Portulaca were inherited from the common ancestor of Portulaca and the cacti.

Because Portulaca species are nested within a predominantly CAM lineage (Guralnick and Jackson, 2001; Nyffeler et al., 2008), it has been previously hypothesized that the C4 pathway of these taxa evolved from an ancestral CAM type (Sage, 2002). This hypothesis is corroborated by the evolutionary history of PEPC genes, with the CAM-specific gene of Portulaca being similar to CAM forms of other species, such as cacti (Fig. 3), while the C4-specific PEPC has been recruited from non-photosynthetic forms. For the other enzymes of the CAM and C4 pathways, Portulaca uses the same genes in both conditions (Supplementary Table S6, available at JXB online). This indicates that, for many enzymes of the CCM, the evolution of one CCM from the other does not require the co-option of new genes. However, since the timing of activity differs between the CCMs, modifications in the regulation of the genes are probably still required. For many of these genes, this may be possible because they are involved with both the C3 pathway and the decarboxylation phase of the C4 pathway that operates during the day in both the CAM and C4 pathways. In the case of PEPC, however, co-option of the CAM-specific gene into the C4 cycle might have been hampered by the distinct regulatory cascades controlling the transcription and translation of the CAM form at night and the C4 form during the day (Jiao and Chollet, 1991; Nimmo, 2003), leading to the recruitment of a distinct gene (namely ppc-1E1a′).

C4 evolution within Portulaca

Portulaca species do not form a homogeneous C4 group, but include a C3–C4 species and several C4 clades that differ in their C4-associated anatomical types and decarboxylating enzymes used for the C4 cycle (Voznesenskaya et al., 2010; Ocampo et al., 2013). In phylogenetic trees of Portulaca species, the C3–C4 taxon is nested within otherwise C4 lineages (Ocampo and Columbus, 2012), which might be interpreted as evidence for a C4 to C3–C4 reversion (Ocampo et al., 2013). However, evolutionary transitions between photosynthetic types are difficult to reconstruct based on species relationships, and C4-related phenotypic and genetic variation can help differentiate alternative scenarios (Christin et al., 2010; Hancock and Edwards, 2014). In the case of Portulaca, the PEPC gene of the C3–C4 P. cryptopetala is nested within those of different C4 clades (Fig. 4). If PEPC had been optimized once for a C4 function at the base of Portulaca, this would have occurred through adaptive evolution on the branch sustaining the whole clade. Such a scenario is ruled out, however, by modelling of codon transitions, which strongly favour a model with adaptive evolution restricted to the branches at the base of each of the three C4 clades (Fig. 4). This shows that the putative C4-specific PEPC of the three C4 clades included in this study underwent adaptive amino acid changes after their divergence, and after their separation from the lineage of P. cryptopetala. For instance, the Ser780 is restricted to genes from the Oleracea clade, while orthologous genes from members of the Pilosa clade underwent other amino acid substitutions that are shared with C4 monocots (e.g. A531P, S761A; Christin et al., 2007). These results show that the optimization of PEPC for a function in C4 photosynthesis occurred independently in each C4 clade, and refutes a C4 to C3–C4 reversal in P. cryptopetala.

In addition to independent optimizations of PEPC genes for C4 photosynthesis, variation exists in the C4-associated anatomy and biochemistry among C4 clades of Portulaca (Voznesenskaya et al., 2010; Ocampo et al., 2013). Based on these observations and our results, the most likely scenario is the addition of a C3–C4 suite of traits over an ancestral
CAM-like type in the common ancestor of *Portulaca*. This C3–C4 type might have been co-opted several times independently for the evolution of a more efficient C4 trait, as suggested for Molluginaceae (Christin et al., 2011). A gradual increase of PEPC activity during the day might then have occurred concomitantly with the development of a more C4-like anatomy, characterized by a high bundle sheath to mesophyll ratio. One way to achieve this state is through high vein densities. Some members of *Portulaca* belong to a handful of lineages in the Portulacineae to have evolved high vein densities via the rearrangement of leaf vasculature into a three-dimensional configuration (Voznesenskaya et al., 2010; Ocampo et al., 2013; Ogburn and Edwards, 2013). While most of these vein rearrangements were associated with large increases in succulence, in *Portulaca* it may have allowed the acquisition of an optimized C4 CCM.

**Conclusions**

Caryophyllales is a hotspot of photosynthetic transitions, with at least 23 C4 and multiple CAM origins. Of three PEPC gene lineages present in eudicots (*ppc-1E1, ppc-1E2*, and *ppc-2*; Fig. 1), only *ppc-1E1* was recurrently recruited into the C4 pathway, suggesting that this gene lineage was more suitable for a C4 function (Christin et al., 2013a). The evidence provided here also supports recruitment of the same gene lineage into CAM metabolism, which suggests that the same capacitated genes present in the common C3 ancestor were co-opted by the numerous CCM origins of Caryophyllales. The evolvability of one CCM compared to the other might depend on the ecology and leaf anatomy of the C3 ancestor of each lineage (Sage, 2002; Edwards and Ogburn, 2012; Edwards and Donoghue, 2013). However, in some cases, evolutionary bridges between the two photosynthetic types exist, as illustrated by *Portulaca*. This shows that while ancestral conditions could influence the evolutionary trajectories of the descendants, the determinism is not perfect and one photosynthetic type can be co-opted to evolve the other.

**Supplementary material**

Supplementary data are available at *JXB* online.

**Figure S1.** Phylogenetic relationships among *ppc-1* genes. This phylogenetic tree was obtained through Bayesian inference on nucleotide sequences. Names of taxonomic groups and gene lineages are indicated on the right. Branches in lineages presenting a Ser780 are highlighted in red. Bayesian support values are indicated near branches. Asterisks indicate putative pseudogenes with one or several stop codons in the coding sequence. Black circles indicate sequences that were isolated from cacti cDNA. (A) Complete phylogenetic tree; (B) *ppc-1E2* of Caryophyllales; (C, D) *ppc-1E1* of Caryophyllales.

**Figure S2.** Phylogenetic relationships among *ppc-2* genes. This phylogenetic tree was obtained through Bayesian inference on nucleotide sequences. Names of taxonomic groups are indicated on the right. Bayesian support values are indicated near branches.

**References**


are accompanied by specific modifications in phosphoenolpyruvate carboxylase. Photosynthesis Research 77, 241–254.


