Key changes in gene expression identified for different stages of C_4 evolution in *Alloteropsis semialata*

Running title: Evolution of C_4 transcriptomes in *Alloteropsis*

Luke T. Dunning[*,1], Jose J. Moreno-Villena[*,1,2], Marjorie R. Lundgren[1,3], Jacqueline Dionora[4], Paolo Salazar[4], Claire Adams[5], Florence Nyirenda[6], Jill K. Olofsson[1], Anthony Mapaura[7], Isla M. Grundy[8], Canisius J. Kayombo[9], Lucy A. Dunning[10], Fabrice Kentatchime[11], Menaka Ariyarathne[12], Deepthi Yakandawala[12], Guillaume Besnard[13], W. Paul Quick[1,4], Andrea Bräutigam[14], Colin P. Osborne[1], Pascal-Antoine Christin[1,a]

[*] These authors contributed equally to this work

[1] Animal and Plant Sciences, University of Sheffield, Western Bank, Sheffield S10 2TN, United Kingdom

[2] Present address: Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT, USA

[3] Present address: Lancaster Environment Centre, Lancaster University, Lancaster LA1 4YQ, United Kingdom[4] International Rice Research Institute, DAPO, Metro Manila, Philippines

[5] Botany Department, Rhodes University, 6140 Grahamstown, South Africa

[6] Department of Biological Sciences, University of Zambia, Lusaka, Zambia

[7] National Herbarium and Botanic Garden, Harare, Zimbabwe

[8] Institute of Environmental Studies, University of Zimbabwe, Harare, Zimbabwe

[9] Forestry Training Institute, Olmotonyi, Tanzania

[10] Department of Social Sciences, University of Sheffield, 219 Portobello, Sheffield S1 4DP, United Kingdom

[11] CABAlliance, P.O. Box 3055 Messa, Yaoundé, Cameroon

[12] Department of Botany, Faculty of Science, University of Peradeniya, Galaha Road, Peradeiya 20400, Sri Lanka

[13] Laboratoire Évolution et Diversité Biologique (EDB UMR5174), Université de Toulouse, CNRS, IRD, UPS, Toulouse, France

[14] Bielefeld University, Universitätsstrasse 35, 33501 Bielefeld, Germany

[a] Corresponding author: Pascal-Antoine Christin; telephone +44-(0)114-222-0027; fax +44 114 222 0002; email: p.christin@sheffield.ac.uk

Date of submission: 26/02/2019

Number of tables: 1

Number of figures: 6

in colour in print: 5 (1, 2, 4, 5, and 6)

in colour online only: none

Word count: 6683

Highlight

Comparative transcriptomics in a phylogenetic context show that the initial emergence of C_4 photosynthesis in *Alloteropsis semialata* coincides with few changes in gene expression within mature leaves, with secondary adaptation occurring in geographically isolated populations.

Abstract

C_4 photosynthesis is a complex trait that boosts productivity in tropical conditions. Compared to C_3 species, the C_4 state seems to require numerous novelties, but species comparisons can be confounded by long divergence times. Here, we exploit the photosynthetic diversity that exists within a single species, the grass *Alloteropsis semialata*, to detect changes in gene expression associated with different photosynthetic phenotypes. Phylogenetically-informed comparative transcriptomics show that intermediates with a weak C_4 cycle are separated from the C_3 phenotype by increases in the expression of 58 genes (0.22% of genes expressed in the leaves), including those encoding just three core C_4 enzymes: ASP-AT, PCK, and PEPC. The subsequent transition to full C_4 physiology was accompanied by increases in another 15 genes (0.06%), including only the core C_4 enzyme PPDK. These changes likely created a rudimentary C_4 physiology, and isolated populations subsequently improved this emerging C_4 physiology, resulting in a patchwork of expression for some C_4-accessory genes. Our work shows how C_4 assembly in *A. semialata* happened in incremental steps, each requiring few alterations over the previous one. These create short bridges across adaptive landscapes that likely facilitated the recurrent origins of C_4 photosynthesis through a gradual process of evolution.

Keywords: adaptation, C_4 photosynthesis, complex trait, intermediates, phylogenetics, transcriptomics

Introduction

The origins of traits composed of multiple anatomical and/or biochemical components have always intrigued evolutionary biologists (Darwin, 1859; Meléndez-Hevia *et al.*, 1996; Lenski *et al.*, 2003). If such traits gain their function only through the co-ordinated action of multiple components, their evolution via natural selection must cross a valley in the adaptive landscape. Despite this obstacle, complex traits have evolved repeatedly in diverse groups of organisms. This apparent paradox is solved for most traits by the existence of intermediate stages, which act as evolutionary enablers, creating bridges over the valleys of the adaptive landscape (Jacob, 1977; Dawkins, 1986; Weinreich *et al.*, 2006; Blount *et al.*, 2012; Vopalensky *et al.*, 2012; Werner *et al.*, 2014). The accessibility of new traits likely depends on the length and complexity of such bridges, which are generally unknown. Quantifying the evolutionary gap between phenotypic states is therefore crucial to contextualise the likelihood of a novel trait evolving.

An excellent system to study the evolutionary trajectories of an adaptive trait is C_4 photosynthesis. This metabolic pathway increases CO_2 concentration at the active site of assimilation via the Calvin-Benson cycle (Hatch, 1987; Sage, 2004; Christin & Osborne, 2014). This avoids the energetically costly process of photorespiration, effectively increasing photosynthetic efficiency in warm and arid conditions (Sage *et al.*, 2012, 2018). This CO_2-concentrating mechanism relies on a set of specific leaf anatomical properties and the co-ordinated action of up to ten enzymes carrying the C_4 reactions (hereafter 'core C_4 enzymes') and numerous associated proteins (Table S1; Hatch, 1987; Bräutigam *et al.*, 2011; Sage *et al.*, 2012; Külahoglu *et al.*, 2014; Lundgren *et al.*, 2014; Yin and Struik 2017). Despite its apparent complexity, C_4 photosynthesis is a textbook example of convergent evolution, having independently evolved more than 60 times within flowering plants (Sage *et al.*, 2011). The origins of C_4 photosynthesis were likely facilitated by the presence of anatomical enablers in some groups (Christin *et al.*, 2013b; Sage *et al.*, 2013), but the processes leading to a functioning C_4 biochemical pathway within these anatomical structures are less well understood. All C_4 enzymes studied so far exist in C_3 plants, but are involved in different pathways (Aubry *et al.*, 2011). There is a bias in the recruitment of genes into the C_4 system, with genes ancestrally abundant in the leaves of C_3 plants preferentially co-opted for C_4 (Christin *et al.*, 2013a; John *et al.* 2014; Emms *et al.*, 2016; Moreno-Villena *et al.*, 2018). Changes to their expression patterns and/or kinetic properties of the encoded enzyme then followed (Bläsing *et al.*, 2000; Hibberd & Covshoff, 2010; Huang *et al.*, 2017; Moreno-Villena *et al.*, 2018), with cell-specific expression realized in some cases through the recruitment of pre-existing regulatory mechanisms (Brown *et al.*, 2011; Kajala *et al.*, 2012; Cao *et al.*, 2016; Reyna-Llorens & Hibberd, 2017; Borba *et al.*, 2018; Reyna-Llorens *et al.*, 2018).

The evolutionary transition between C_3 and C_4 phenotypes involves intermediate stages that only have some of the anatomical and biochemical modifications typical of C_4 plants (Monson *et al.*, 1989; Sage *et al.*, 2012, 2018). In particular, some C_3+C_4 plants perform a weak C_4 cycle that is responsible for only part of their carbon assimilation (these correspond to 'type II C_3-C_4 intermediates'; Ku *et al.*, 1983; Monson *et al.*, 1986; Schlüter & Weber, 2016). This weak C_4 cycle might have emerged through the upregulation of C_4-related enzymes to balance nitrogen among cellular compartments in the multiple lineages of plants that use a photorespiratory pump (Sage *et al.*, 2011, 2012; Mallmann *et al.*, 2014; Bräutigam & Gowik, 2016). Metabolic models suggest that any increase in flux of CO_2 fixed through the C_4 cycle in intermediate plants directly translates into biomass gain, leading to gradual increases in C_4 gene expression (Heckmann *et al.*, 2013; Mallmann *et al.*, 2014). The current model of C_4 evolution therefore assumes gradual, yet abundant changes in plant transcriptomes and genomes during the transition from C_3 ancestors to physiologically C_4 descendants. Indeed, comparisons of C_3 and C_4 species have typically identified thousands of differentially expressed genes encoding C_4 enzymes, regulators, and accessory metabolite transporters (Bräutigam *et al.*, 2011, 2014; Gowik *et al.*, 2011; Külahoglu *et al.*, 2014; Li *et al.*, 2015; Lauterbach *et al.*, 2017). These large numbers might partially result from the comparison of species typically separated by millions of years of divergence (Christin *et al.*, 2011), which leaves ample time for the accumulation of secondary changes linked to the C_4 trait beyond the minimal requirements, as well as variation in other unrelated traits (Heyduk *et al.* In press). Even within a single species where photosynthetic transitions can be induced, the number of differentially expressed genes identified in transcriptome comparisons can be extremely high (Chen *et al.*, 2014). Previous efforts have however typically targeted very few individuals per C_4 lineage, such that the initial bout of co-option that generated a C_4 cycle cannot be distinguished from subsequent adaptation via natural selection and diversification caused by genetic drift (Christin & Osborne, 2014; Reeves *et al.*, 2018; Heyduk *et al.* In press).

In this study, the transcriptomes of mature leaves are compared among plant populations using a phylogenetic approach. The work aims to quantify the phenotypic differences in gene expression between the C_3 phenotype and plants using a weak C_4 cycle (C_3+C_4 state), independently from those responsible for the transition to the full C_4 type, and finally from those involved in the adaptation of an existing C_4 phenotype. The time elapsed between transitions, and therefore the number of changes unrelated to C_4 emergence, is reduced by focusing on a single species containing a diversity of photosynthetic types, the grass *Alloteropsis semialata*. Congeners of *A. semialata* are C_4, but previous comparative transcriptomics and leaf anatomy have shown that C_4 biochemistry emerged multiple times in the genus, from a common ancestor with some C_4-like characters (Fig. 1; Dunning *et al.*, 2017). Capitalizing on the physiological diversity existing within *A. semialata*, leaf transcriptomes from multiple individuals originating from diverse populations of each photosynthetic type in this species are analysed, together with closely related C_3 and C_4 species, to detect the changes in gene expression linked to (i) the phenotypic difference between C_3 plants and C_3+C_4 intermediates, (ii) the shift to fixing carbon exclusively via the C_4 pathway in solely C_4 plants, and (iii) the adaptation of the C_4 cycle after its evolution in geographically isolated C_4 populations. This deconstruction of the genetic origins of a complex biochemical pathway sheds new light on the number of genetic changes needed to move to another part of the adaptive landscape during different stages of a stepwise physiological transition.

Material and Methods

Species sampling and growth conditions

Three biological replicates from ten separate populations/species were used for differential gene expression analyses. Seven of these were geographically distinct *Alloteropsis semialata* populations including: two C_3 populations from South Africa (RSA6) and Zimbabwe (ZIM1502) that represent extremes of the C_3 geographic range (Fig. 1B; Lundgren *et al.*, 2015), two geographically distant C_3+C_4 populations from Tanzania (TAN1602) and Zambia (ZAM1503) that are hypothesised to operate a weak C_4 cycle (Lundgren *et al.*, 2016), and three C_4 populations from Cameroon (CMR1601), Tanzania (TAN4) and the Philippines (PHI1601) that sample the two C_4 genetic subgroups (Olofsson *et al.*, 2016; Fig. S1). The C_4 populations of *A. semialata* have decreased CO_2-compensation points, increased carboxylation efficiencies, and shifts in carbon isotopes compared with the C_3 populations that confirm their photosynthetic type (Lundgren *et al.*, 2016). The C_4 leaves are characterized by increased vein density, PEPC protein abundance, and transcript abundance of genes encoding some C_4 enzymes compared with the C_3 types (Lundgren *et al.*, 2016, 2019; Dunning *et al.*, 2017). The C_3+C_4 *A. semialata* also show elevated leaf levels of PEPC protein and genes for some C_4 enzymes and increased concentration of chloroplasts in bundle sheaths in comparison with the C_3 populations, but no increase in vein density (Lundgren *et al.*, 2016; Dunning *et al.*, 2017). However, while slightly shifted compared to their C_3 conspecifics, their carbon isotope ratios are not in the C_4 range, which is common in plants performing a weak C_4 cycle, responsible for only part of their CO_2 uptake (i.e. ‘type II intermediates’; Monson *et al.*, 1988; von Caemmerer, 1992; Sage *et al.*, 2012; Lundgren *et al.*, 2016). This results in a reduced CO_2-compensation point and oxygen inhibition (Lundgren *et al.*, 2016), as observed in other species acquiring part of their carbon via a weak C_4 cycle (Ku *et al.*, 1991). In addition to the seven *A. semialata* populations, we included one population of each of the C_4 congeners *A. angusta* (AANG1 from Uganda) and *A. cimicina* (from Madagascar) to enable comparison of convergent C_4-related changes in gene expression (Fig. S1). Finally, an *Entolasia marginata* population from Australia was included as a C_3 outgroup. Three distinct genotypes for eight of the ten populations described above were retrieved from a recent dataset (Dunning *et al.* In press) or sequenced here. For the two other populations, sufficient biological replicates were not available. For *A. angusta*, we sequenced three clones of a single wild collected plant that were established more than one year before the study, while for *E. marginata* we sequenced two different genotypes and a clone of one of these genotypes, similarly established before the study (See Table S2 for detailed sample collection information).

To evaluate the diversity of gene expression across the diversity of photosynthetic types and the genetic diversity within each photosynthetic type, we supplemented the above data with a single biological replicate from a further 15 geographic distinct populations (12 from previously published data; Dunning *et al.*, 2017, In press; Fig. 1A). The three newly sequenced individuals are two C_4 *A. semialata* from Sri Lanka (SRI1702, lat: 6.81 long: 80.92) and Zambia (ZAM1726, lat: -14.21 long: 28.60), and a C_3 individual from Zimbabwe (ZIM1503, lat: -18.78 long: 32.74). In total, we had 45 RNA-Seq libraries from 25 populations/species, with three biological replicates sampled from 10 populations and a single biological replicate sampled from the remaining 15 populations (Fig. 1A).

All plants were collected from the field as seeds or live cuttings, and subsequently grown under controlled conditions at the University of Sheffield as previously described (Dunning *et al.*, 2017). In brief, plants were potted in John Innes No. 2 compost (John Innes Manufacturers Association, Reading, England) and maintained under wet, nutrient-rich conditions in controlled environment chambers (Conviron BDR16; Manitoba, Canada) set to 60% relative humidity, 500 μmol m^{-2} s^{-1} light intensity, 14h photoperiod, and day/night temperatures of 25/20°C. After a minimum of 30 days in these growth conditions, young fully expanded leaves were sampled for transcriptome analyses.

RNA extraction, sequencing, and transcriptome assembly

RNA extraction, library preparation and sequencing were performed as previously described (Dunning *et al.*, 2017). In brief, total RNA was extracted from the distal half of fully expanded fresh leaves, sampled in the middle of the light period, using the RNeasy Plant Mini Kit (Qiagen, Hilden, Germany) with an on-column DNA digestion step (RNase-Free Dnase Set; Qiagen, Hilden, Germany). Total RNA was used to generate 34 indexed RNA-seq libraries using the TruSeq RNA Library Preparation Kit v2 (Illumina, San Diego, CA). Each library was subsequently sequenced on 1/24 of a single Illumina

HiSeq 2500 flow-cell (with other samples from the same or unrelated projects), which ran for 108 cycles in rapid mode at the Sheffield Diagnostic Genetics Service.

The raw RNA-Seq data were cleaned using the Agalma pipeline v.0.5.0 to remove low quality reads (Q<30), and sequences corresponding to ribosomal RNA or containing adaptor contamination (Dunn *et al.*, 2013). *De novo* transcriptomes were assembled using Trinity (version trinityrnaseq_r20140413p1; Grabherr *et al.*, 2011). All raw data and transcriptome assemblies have been submitted to the NCBI repository (Bioproject PRJNA401220). Coding sequences (CDS) longer than 500 bp were predicted for each population using OrfPredictor (Min *et al.*, 2005), which uses homology to a user supplied reference protein database or *ab initio* predictions if no suitable match is found. The protein database used comprised the complete coding sequences of eight model species: *Arabidopsis thaliana*, *Brachypodium distachyon*, *Glycine max*, *Oryza sativa*, *Populus trichocarpa*, *Setaria italica*, *Sorghum bicolor* and *Zea mays.*

Phylogenetic reconstruction using core-orthologs

Single-copy orthologs were extracted from the newly and previously published transcriptome assemblies (Dunning *et al.*, 2017) to infer phylogenetic relationships among individuals. Homologous sequences to 581 single-copy plant core-orthologs previously determined in the Inparanoid ortholog database (Sonnhammer & Ostlund, 2014) were identified. A Hidden Markov Model based search tool (HaMSTR v.13.2.3; Ebersberger *et al.*, 2009) was used to screen the CDS of the transcriptomes. Sequences of the single copy plant core-orthologs were subsequently aligned using a previously described stringent alignment and filtering pipeline (Dunning *et al.*, 2017). In brief, the CDS were translation aligned and filtered using T-COFFEE v. 11.00.8cbe486 (Notredame *et al.*, 2000) before trimming with gblocks v.0.91 (Castresana, 2000). Sequences shorter than 100 bp after trimming, and ortholog alignments with a mean nucleotide identity <95% were discarded, retaining 504 markers. A maximum likelihood tree was inferred using IQ-TREE v.1.6.3 (Nguyen *et al.*, 2014), which determined the most appropriate nucleotide substitution model prior to inferring a phylogeny with 1,000 ultrafast bootstrap replicates.

Differential expression analyses

For differential expression analysis, we used the 45,144 cDNA sequences from the *A. semialata* reference genome (Dunning *et al.*, In press; accession number QPGU00000000) as a reference. Cleaned reads were mapped to the reference using Bowtie2 v.2.3.4.1 (Langmead & Salzberg, 2012) recording all alignments. Counts for each transcript were then calculated using eXpress v.1.5.1

(Roberts & Pachter, 2013) with default parameters, and are reported in reads per kilobase of transcript per million mapped reads (rpkm). A multivariate analysis was used to assess similarities and differences in overall transcriptome expression profiles between samples. Clustering of expression profiles based on the biological coefficient of variation (BCV) were identified with multidimensional-scaling (MDS) in edgeR v3.4.2 (Robinson *et al.*, 2010).

Differential expression analysis in edgeR was restricted to the ten populations with three biological replicates. For each pair of populations, differentially expressed genes were identified as those with an associated false discovery rate (FDR) below 0.05. The overlap between pairwise comparisons was used to identify changes associated with specific branches of the phylogenetic tree inferred from core orthologs. Changes were assigned to a branch if significant results were detected for all pairwise tests involving one member of the descending clade and one population outside the clade, and the direction of expression change was consistent. This summary of pairwise tests was done separately for each C_3+C_4/C_4 clade (*A. cimicina*, *A. angusta*, and *A. semialata*) with all C_3 populations so that convergent gene expression shifts could be detected. Overall, by grouping the differential expression results based on the phylogenetic clades, we are able to identify changes in gene expression that coincide with specific physiological transitions, as well as those that precede or follow these transitions.

Results

Transcriptome sequencing

Over 190 million 108-bp paired-end reads were used in this study, including more than 167 million for the ten populations sampled in triplicate (Table S3). For these 30 samples used in differential expression analyses, the data comprised 36.13 Gb, with a mean of 1.20 Gb per library (SD=0.54 Gb; Table S3). Over 95% of reads were retained after cleaning, and a *de novo* transcriptome was assembled for each of the populations using all available reads.

Phylogenetic relationships based on concatenated ortholog alignments

A phylogenetic tree was inferred from a concatenated alignment of 504 'core-orthologs' extracted from the predicted coding sequences from 25 transcriptome assemblies (12 assembled here), for a total of 573,762 bp after cleaning. Each population was represented by at least 126,048 bp (mean=468,507 bp; SD 94,782 bp). The concatenated alignment had 21.1% gaps and 6.3% of sites were parsimony informative. The phylogeny was inferred using the GTR+F+R4 substitution model, which was the best fit model according to the BIC. The phylogenetic relationships were congruent with previous genome-wide nuclear trees (Olofsson *et al.*, 2016; Dunning *et al.* In press), and confirmed that all the sampled C_4 populations of *A. semialata* form a monophyletic group, which is sister to the C_3+C_4 populations (Fig. 1). These two are in turn sister to the C_3 populations, so that previously inferred nuclear clades I (C_3), II (C_3+C_4), III and IV (both C_4) are retrieved, with the polyploid populations (RSA3 and RSA4) branching in between and the Cameroonian population at their base (Olofsson *et al.*, 2016; Fig. 1). *A. angusta* and *A. cimicina* branched successively outside of *A. semialata* (Fig. 1), again mirroring previous results (Lundgren *et al.*, 2015; Olofsson *et al.*, 2016; Dunning *et al.* In press).

Transcriptome-wide patterns

A mean of 57.4% (SD=12.05%) of cleaned reads from the 45 RNA-Seq libraries mapped back to the 45,144 cDNA sequences extracted from the reference *A. semialata* genome (only *A. semialata* samples n=34, mean=64.1%, SD=4.3%). In total, 59.8% (n=26,975) of gene sequences had expression levels of >1 read per million of mapped reads in at least three samples and were retained for differential expression analysis. Based on their expression profiles, samples group strongly by species (Fig. 2A). When focusing on *A. semialata*, the main phylogenetic groups are recovered, which match the photosynthetic types (Fig. 1 and 2B). There is no apparent effect of the source study, with previous and new transcriptomes of the same species grouping together (Fig. 2). Differential expression analysis was performed for each pair of the ten populations that had three biological replicates. The 45 pairwise tests performed returned an average of 4,880 (SD=2,125) significantly (FDR<0.05) differentially expressed genes (Fig. 3; Table S4). The number of differentially expressed genes is highest between the most distantly-related populations and lowest among close relatives (Fig. 3). Complete expression results are available in Tables S4 and S5.

Differences between the C_3 and C_3+C_4 states of A. semialata

As expected, the long divergence time between the C_3 outgroup (*Entolasia marginata*) and *A. semialata* results in a large number of significant expression changes (branch A in Fig. 4). A total of 825 genes are downregulated along this branch (3.1% of those expressed in leaves), including two genes encoding phosphoenolpyruvate carboxylase (PEPC; *ppc-1P2* and *ppc-2P1*; ASEM_AUS1_43423 and ASEM_AUS1_37421; Table S6), which drop to barely detectable levels in all *A. semialata* accessions, and are therefore unlikely to be linked to photosynthetic diversification. A total of 1,500 genes (5.6%) are upregulated in *A. semialata* compared to the C_3 outgroup (branch A in Fig. 4; Table S6). This includes genes encoding the C_4-related enzymes malate dehydrogenase (NAD-

MDH; *nadmdh-2P4*; ASEM_AUS1_14800), adenosine monophosphate kinase (AK; ak-3P3; ASEM_AUS1_08191 and ASEM_AUS1_08195), glyceraldehyde 3-phosphate dehydrogenase (GAPDH; *gapdh-1P2;* ASEM_AUS1_06811) and phosphoenolpyruvate carboxylase kinase (PEPC-K; *pepck-1P3* and *pepck-3P6;* ASEM_AUS1_38337 and ASEM_AUS1_12272), although their expression levels remain fairly low in all *A. semialata* regardless of photosynthetic type (mean=42 rpkm; SD=37; Table S5). One gene encoding an enzyme linked to the photorespiratory pathway is also upregulated (*hpr-2P3;* ASEM_AUS1_28984), although levels again remain fairly low within *A. semialata* (mean=19 rpkm; SD=13; Table S5). The rest of the numerous genes varying in expression between the whole of *A. semialata* and the outgroup do not have known links to the C_4 pathway. A total of 60 genes (0.22%) are differentially expressed along the branch leading to the C_3 populations of *A. semialata* (branch B in Fig. 4). None of these 60 genes encodes a protein known to function as part of the C_4 pathway (Table S6).

Within *A. semialata*, a C_4 cycle, weak or strong, characterizes the monophyletic group of C_3+C_4 and C_4 populations, but not its C_3 sister group. Along the branch leading to C_3+C_4 and C_4 accessions we detect 67 significantly differentially expressed genes (branch E in Fig. 4; Table 1). Of those, 58 (0.22% of all expressed genes) are consistently upregulated in the C_3+C_4 and C_4 populations compared to the C_3 samples, including three genes that encode key C_4 enzymes: aspartate aminotransferase (ASP-AT; *aspat-3P4;* ASEM_AUS1_08268), phosphoenolpyruvate carboxykinase (PCK; *pck-1P1;* ASEM_C4_17510), and PEPC (*ppc-1P3;* ASEM_C4_19029; Table S6). These three genes reach very high levels in the leaves of all C_3+C_4 and C_4 individuals (mean=1,766 rpkm; SD=585; Table S5; Fig. 5), including the C_4 congener *A. angusta* (mean=5,002 rpkm; SD=2,607; Table S5). The other genes whose expression changes significantly along the same branch mostly remain at low to moderate levels in all *A. semialata*, but a number of them are also significant in *A. angusta*, and for two of them in *A. cimicina* (Tables 1 and S6). The significant genes include one for Nudix hydrolase, which was previously identified in a comparison of rice and C_4 grasses (Ding *et al.*, 2016). The remaining genes have however not been related to C_4 photosynthesis in previous screens of grasses (Ding *et al.*, 2016; Huang *et al.*, 2017). A gene for a callose synthase is downregulated in the C_3+C_4/C_4 group as well as *A. angusta* (Table 1), which might be linked to plasmodesmatal widening to facilitate intercellular fluxes, as suggested for other genes linked to callose synthesis (Bräutigam *et al.*, 2011; Huang & Brutnell, 2016). Some of the other differentially expressed genes encode proteins that have been previously suggested as being involved in metabolic/structural differences between photosynthetic types (e.g. acyl transferase, pyruvate dehydrogenase; Huang & Brutnell, 2016) or that might be linked to plasmodesmata (e.g. phosphatidylglycerol/phosphatidylinositol transfer protein), although the functional links with photosynthetic diversification remain to be tested.

Changes during the transition from C_3+C_4 to C_4 in A. semialata

Within *A. semialata*, a strong C_4 cycle characterizes a monophyletic group of populations (Fig. 1A), but only 16 genes (0.06% of all expressed genes) were significantly differentially expressed along the branch separating this group from the other populations (branch I in Fig. 4). Of these, 15 were consistently upregulated in the C_4 populations, including one gene encoding the core C_4 enzyme pyruvate orthophosphate dikinase (PPDK; *ppdk-1P2;* ASEM_AUS1_39556), which reaches very high levels in all C_4 populations (mean=4,479 rpkm; SD=2,293; Tables 1 and S6; Fig. 5), including the congeners *A. cimicina* (mean=1,766 rpkm; SD=585; Table S5) and *A. angusta (*mean=1,367 rpkm; SD=1,100; Table S5). The other genes upregulated in the C_4 accessions, which include transcription factors and some transporters, reach moderate levels in the C_4 accessions, although some are also significantly upregulated in *A. angusta* (Table 1). Significant changes in the abundance of the genes for the phosphatidylglycerol/phosphatidylinositol transfer protein might be linked to modifications of plasmodesmata to facilitate metabolite exchanges (Grison *et al.*, 2015), while aquaporins might be involved in membrane diffusion of CO_2 (Kaldenhoff *et al.*, 2014). However, whether these genes played a direct role in the photosynthetic diversification of *A. semialata* remains speculative.

Adaptation of C_4 photosynthesis in independent lineages

The three C_4 populations included in the differential expression analyses come from geographically distant locations and diverged more than half a million years ago (Lundgren *et al.*, 2015; Olofsson *et al.*, 2016), explaining the large number of differentially expressed genes among them (Fig. 3). Interestingly, this includes enzymes linked to the C_4 cycle with genes encoding PEPC (*ppc-1P3*; ASEM_AUS1_12633), NAD-MDH (*nadmdh-1P8;* ASEM_AUS1_25602), PEPC-K (*pepck-1P3;* ASEM_C4_38337), NADP-MDH (nadpmdh-3P4; ASEM_AUS1_33376), and a sodium bile acid symporter (SBAS; *sbas-4P4;* ASEM_AUS1_12098) all upregulated in the C_4 plants from the Philippines (PHI1601; Table S6). A comparison of expression levels in the other transcriptomes (including the 15 populations not used for the differential expression) indicates that the gene *sbas-4P4* has qualitatively higher expression in all C_4 individuals from clade IV of *A. semialata* (mean=898 rpkm; SD=483), but not in the other C_4 individuals (mean=27 rpkm; SD=19) or the other *A. semialata* populations as a whole (mean=20 rpkm; SD=13; Table S5; Fig. 5). This gene is orthologous to a group of *Arabidopsis* paralogs including BASS6 (At4g22840), which has the ability to transport glycolate, and appears to be involved in a process decreasing photorespiration (South *et al.*, 2017). The

Arabidopsis paralog previously related to C_4 photosynthesis transports pyruvate (BASS2; Furumoto *et al.*, 2011), but its precise function might differ between the *Alloteropsis* and *Arabidopsis* orthologs. In addition, a gene encoding the photorespiratory enzyme peroxisomal (S)-2-hydroxy-acid oxidase (GLO; *glo*-1P1; ASEM_AUS1_30871) is downregulated in only one of the three C_4 populations (CMR1601; Table S6).

There is quite large variation in the expression of individual genes encoding some other C_4 enzymes, with some more abundant in the C_4 than C_3+C_4 *A. semialata* populations on average, yet relatively low in other C_4 individuals. These genes include alanine aminotransferase (ALA-AT; alaat-1P5, ASEM_AUS1_25403; C_4 mean=1,105 rpkm; SD=812; C_3+C_4 mean=134 rpkm; SD=59; significantly differentially expressed in 13 of the 15 required pair-wise tests), which has low expression in C_4 individuals from Tanzania (TAN4-08; rpkm=135) and Cameroon (CMR1601-07; rpkm=154). Similarly, one of the genes encoding the NADP-malic enzyme (*nadpme-1P4*; NADP-ME, ASEM_AUS1_06611; significantly differentially expressed in 7 of the 15 required pair-wise tests) is on average more abundant in the C_4 and C_3+C_4 (mean=300 rpkm; SD=235) than C_3 (mean=75 rpkm; SD=32) *A. semialata* populations, but low within some C_4 individuals (e.g. TAN4-01 rpkm=82; TAN4-08 rpkm=54; ZAM1503-08 rpkm=50; Fig. 5). This gene is also significantly upregulated in *A. cimicina* and *A. angusta* (Table S5). One of the genes for PEPC kinase (*pepck1P3*) reaches high levels in several C_4 accessions of *A. semialata* (Table S5). Similarly, some genes for the small unit of Rubisco reach very low levels in some C_4 accessions. For instance, the gene AUS1_20231 is at low levels in most C_4 *A. semialata*, yet remains very high in others while the paralog AUS1_26631 reaches extremely low levels, specifically in the Asian group of C_4 *A. semialata* (Table S5). A third paralog (AUS1_26630) remains high in all accessions, so that the total abundance of genes for Rubisco is not markedly decreased, which is congruent with the high Rubisco protein abundance in the leaf of the C_4 *A. semialata* (Ueno & Sentoku, 2006).

The number of genes significantly differentially expressed in the C_4 *A. cimicina* and *A. angusta* lineages is much higher, since only one population represents each of these species (Fig. S3). As previously reported (Dunning *et al.*, 2017), a high number of genes encoding core C_4 enzymes, regulatory proteins and transporters are upregulated in *A. cimicina* (Table S7), and to a lesser extent in *A. angusta* (Table S8), while some photorespiration and Rubisco genes are downregulated in both species. Besides the differentially expressed genes, a number of C_4-related genes are abundant in all samples independent of their photosynthetic type. This is especially the case of genes encoding β-carbonic anhydrase (*βca-2P3*; ASEM_AUS1_16750; mean=1,682 rpkm, SD=1,027, min=290) and malate dehydrogenases: *nadpmdh-1P1* (ASEM_AUS1_23802; mean=443 rpkm, SD=501, min=117), *nadpmdh-3P4* (ASEM_AUS1_33376; mean=447 rpkm, SD=184, min=166), and *nadmdh-3P5* (ASEM_AUS1_22160; mean=157 rpkm, SD=69, min=41). Transcripts for these genes were also abundant in the leaves of distantly related C_3 grasses, and their upregulation very likely predates the diversification of the group (Moreno-Villena *et al.*, 2018).

Discussion

Sampling the natural diversity to limit false positives

RNA-Seq is routinely used to identify genes differentially expressed between individuals with distinct phenotypes, leading to lists of candidate genes underpinning these differences (e.g. Shen *et al.*, 2014; Dunning *et al.*, 2016; Fracasso *et al.*, 2016). When comparing distinct species, the risk of false positives is very high, as all changes in gene expression unrelated to the studied phenotypic transitions are detected. Here, 77.1% of genes expressed in the leaves are significantly differentially expressed in at least one pairwise comparison between our ten populations (49.8% within *A. semialata*), which all belong to a relatively small group of closely related grasses. A powerful strategy to reduce false positives is to consider multiple independent origins of the trait of interest, and retain only those genes differentially expressed in all lineages (Ding *et al.*, 2016; Rao *et al.*, 2016). Such a filter would however exclude non-convergent changes in gene expression.

The alternative approach adopted here was to carry out multi-individual comparisons to infer changes along specific branches of the phylogenetic tree. The problem of false positives remains, as changes coinciding with the studied transitions would also be detected. However, working within a species complex decreases the number of false positives, as shorter divergence times are likely to result in fewer unrelated changes in gene expression. Because most changes cluster on terminal branches (Fig. 4), probably representing neutral changes that do not persist over evolutionary time, the inference of changes on short internal branches is less likely to be affected by drift. Indeed, a comparison of a C_3 *A. semialata* with the C_4 sister species *A. angusta* would identify over 5,000 (18% of genes expressed in the leaves) differentially expressed genes (Fig. 3). This number drops by approximately 50% when comparing individual C_3 and C_4 populations within *A. semialata*, but still includes all changes that occurred before, during, and after the C_3 to C_4 transition. After incorporating multiple populations of each type, only 67 genes (0.25% of genes expressed in the leaves) are identified that differ in expression between the C_3 and C_3+C_4 phenotypes, and 16 (0.06% of genes expressed in the leaves) between the C_3+C_4 and C_4 states. Changes in some of these genes might not be directly linked to the diversification of photosynthetic types, but several were convergently modified in *A. angusta* and/or *A.*

cimicina (Table 1). These genes represent the best candidates for a role in the emergence and subsequent strengthening of a C_4 cycle in the group.

Emergence and reinforcement of the C_4 cycle in Alloteropsis semialata

The phylogenetic relationships and genus-wide comparisons of transcriptomes and leaf anatomical traits indicate that the last common ancestor of all *A. semialata* might have possessed a weak C_4 cycle based on the upregulation of some enzymes (Fig. 1; Dunning *et al.*, 2017). A large number of genes are differentially expressed between all *A. semialata* and the C_3 outgroup, which is not surprising given the evolutionary distance of at least 15 Myr (Christin *et al.,* 2014). However, these include relatively few genes encoding C_4 enzymes (Table S6). We conclude that the transcriptome of the C_3 *A. semialata* differs from that of other C_3 grasses by relatively few C_4-related genes. The C_3 group might represent a reversal from a C_3+C_4 state to a phenotype with expression levels similar to the C_3 outgroup. In such a scenario, C_4-related changes that happened in the last common ancestor of *A. semialata* and were reversed in the C_3 group would be assigned to the branch leading to the C_3+C_4 and C_4 groups. Because they focus on the phenotypic gaps in gene expression between the C_3 state and those using a weak or strong C_4 cycle, our transcriptome comparisons are therefore not heavily influenced by potential evolutionary reversals or reticulate evolution.

In total, 67 genes are differentially expressed in the group encompassing C_3+C_4 and C_4 phenotypes, and these include only three genes encoding core C_4 enzymes that are upregulated in all C_3+C_4 and C_4 individuals (genes for ASP-AT, PCK and PEPC; Table 1; Table S5). These three enzymes form an aspartate shuttle based on the PCK decarboxylase (Fig. 6), which theoretically cannot sustain a full C_4 pathway on its own without creating an energetic imbalance among cell types (Wang *et al.*, 2014). However, it might create a weak CO_2-concentrating mechanism in C_3+C_4 plants that can function without dramatic energetic consequences due to its coexistence with a C_3 type of photosynthesis. While the functional significance of the other changes detected along the same branch is not always known, several might be linked to the control of plasmodesmata and thereby intracellular exchanges (Table 1). Other small adjustments of the cellular metabolism might remain undetected, but none of the other major C_4 enzymes or transporters are significantly upregulated during the emergence of a weak C_4 cycle (Table 1). The apparently few changes in transcription required to operate a weak C_4 cycle in the C_3+C_4 intermediates may be facilitated by C_4-like anatomical properties and an abundance of genes for some key enzymes in the ancestor, as observed in other C_3 grasses (Christin *et al.*, 2013a, 2013b; Emms *et al.*, 2016; Dunning *et al.*, 2017; Moreno-Villena *et al.*, 2018), and recent evidence suggests that some anatomical traits themselves might emerge via very few genetic changes (Wang *et* *al.*, 2017). While it is only responsible for part of the plant's CO_2 uptake, the weak C_4 cycle of C_3+C_4 plants reduces photorespiration (Ku *et al.*, 1991; Lundgren *et al.*, 2016), which confers a selective advantage analogous to that of a complete C_4 cycle in tropical conditions (Sage *et al.*, 2012; Christin & Osborne, 2014; Lundgren & Christin, 2017), and allows the evolution of a stronger C_4 cycle under natural selection for faster biomass accumulation (Heckmann *et al.*, 2013; Mallmann *et al.*, 2014; Bräutigam & Gowik, 2016).

The transition from a weak to a strong C_4 cycle in *A. semialata* changes carbon isotope signatures (the method most often used to identify photosynthetic types) from non-C_4 values to values diagnostic of C_4 plants (von Caemmerer, 1992; Lundgren *et al.*, 2015). This shift indicates a strengthened connection between the C_3 and C_4 cycles and a decreased leakiness, so that less atmospheric CO_2 is directly fixed by the Calvin-Benson cycle (Monson *et al.*, 1988; von Caemmerer, 1992). Within *A. semialata*, this might have been mediated by the reduced distance between veins in the C_4 *A. semialata* (Lundgren *et al.*, 2016, 2019; Dunning *et al.*, 2017) and/or biochemical alterations. The upregulation of relatively few genes (0.06%) coincided with the phenotypic transitions, and only one of these encoded an enzyme with a known C_4 function, namely PPDK. This enzyme is responsible for the regeneration of PEP, the substrate of PEPC (Fig. 6). An increased PPDK activity is also observed between species of *Flaveria* performing a weak and a strong cycle, and it has been suggested that this provides PEPC with PEP at higher rates, thereby increasing the efficiency of the C_4 pathway (Monson & Moore, 1989; Sage *et al.*, 2012). Based on the literature and our transcriptome data, the C_4 cycle of *A. semialata* relies on a minimum of seven enzymes (Fig. 6; Frean *et al.*, 1983; Ueno & Sentoku, 2006). Genes for some of these enzymes (NAD-MDH, and AK) increased in the common ancestor of the whole group, potentially as part of an ancestral weak C_4 cycle (Fig. 1; Dunning *et al.*, 2017). Within *A. semialata*, further increases in transcript abundance are observed in the C_3+C_4 vs C_3 or C_4 vs C_3+C_4 comparisons (Table 1) for genes encoding PEPC and three other enzymes (i.e. ASP-AT, PCK, and PPDK; Fig. 5). The expression of genes encoding CA and others NAD(P)-MDH in the C_3 ancestor of the group might have been sufficient to sustain a functioning C_4 cycle (Table S5; Moreno-Villena *et al.*, 2018). Genes for the last of these enzymes (NADP-ME) are abundant in some C_4 individuals (Table S5; Fig. 5), and might be expressed only in specific conditions, as suggested previously (Frean *et al.*, 1983).

C_4 populations of *A. semialata* are also characterized by a set of specific anatomical modifications and changes in the cellular localization of some enzymes (Ueno & Sentoku, 2006; Lundgren *et al.*, 2016, 2019; Dunning *et al.*, 2017). Gene expression changes responsible for these modifications would not necessarily be captured by our transcriptome analyses of full mature leaves, and the evolution of the C_4 phenotype almost certainly involves more genetic changes than those detected here. While protein abundance is not a direct function of gene expression, the two are correlated (Schwanhäusser *et al.*, 2011; Csárdi *et al.*, 2015; Koussounadis *et al.*, 2015). In the case of *A. semialata*, the three C_4 enzymes with genes differentially expressed in the C_3+C_4/C_4 transcriptomes (PEPC, ASP-AT and PCK) are also the ones with large differences in activities between the C_3 and C_4 *A. semialata* in a previous study (Ueno & Sentoku, 2006). Transcriptome comparisons offer a first assessment of the changes underlying adaptive transitions, allowing subsequent investigations of responsible regulatory elements, post-transcriptional processes, changes of the protein kinetics, and verification of gene functions via genetic manipulation (e.g. Wang *et al.*, 2017; Borba *et al.*, 2018). Overall, our comparative transcriptomics show that, once the required enablers are present, the transition between C_3 to C_3+C_4 with some C_4 activity, and C_3+C_4 to a rudimentary C_4 metabolism might have required fewer changes in gene expression in *A. semialata* than previously suggested based on other comparisons (Bräutigam *et al.,* 2011, 2014; Gowik *et al.,* 2011; Külahoglu *et al.,* 2014; Li *et al.,* 2015). These changes were spread between the C_3/C_3+C_4 and C_3+C_4/C_4 transitions, supporting a stepwise model of evolution (Mallmann *et al.,* 2014), where evolutionarily stable adaptive peaks can be reached with few mutations.

Adaptation continued after the emergence of a rudimentary C_4 pathway

The CO_2-pump generated by the C_4 cycle of *A. semialata* is less efficient than that of other C_4 species (Niklaus and Kelly, 2019), as illustrated by the incomplete segregation of enzymes between different cell types (Ueno & Sentoku, 2006) and slightly elevated CO_2-compensation points lying at the upper limit of those observed in C_4 species (Lundgren *et al.,* 2016). Therefore, *A. semialata* may be considered to exhibit an incipient C_4 cycle, which has not been optimised through protracted evolutionary periods, as suggested in the most recent models (Bräutigam & Gowik, 2016). The analyses conducted here, which compared all C_4 individuals to the C_3+C_4 or C_3 conspecifics, can detect the changes that happened in the early C_4 members of the group, before the diversification of the C_4 genotypes. However, transcriptome comparisons across C_4 individuals of *A. semialata* show evidence of additional alterations of the leaf biochemistry subsequent to the initial emergence of a C_4 cycle, with the abundance of some C_4-related enzymes varying in abundance across C_4 populations (e.g. NAD-MDH) and photorespiratory proteins downregulated in only some of the C_4 populations (Tables S5 and S6). These changes likely represent the adaptation of the C_4 cycle after its initial emergence (Heyduk *et al.* In press; Niklaus and Kelly, 2019), previously illustrated for *A. semialata* by variation in the identity of genes responsible for an abundance of the key C_4 enzyme PEPC across C_4 genotypes (Dunning *et* *al.*, 2017) and leaf anatomy (Lundgren *et al.* 2019), and recently reported for *Gynandropsis gynandra* (Reeves *et al.*, 2018).

The C_4 pathway proposed for *A. semialata*, based on the upregulation of four core C_4 enzymes in addition to those present in C_3 ancestors (Fig. 6), might serve as an intermediate stage toward more complex and more efficient C_4 cycles. The congeneric C_4 *A. cimicina* and *A. angusta* have transcriptomes more typical of other C_4 species, with very high levels of numerous C_4-related enzymes, including a number of regulatory proteins and metabolite transporters (Table S5), as would be predicted from other study systems, and an abundance of amino acid transitions adapting the proteins for the new catalytic context (Bräutigam *et al.*, 2011, 2014; Gowik *et al.*, 2011; Mallmann *et al.*, 2014; Christin *et al.*, 2015; Dunning *et al.*, 2017). These two species might have undergone more adaptive changes, due to an earlier C_4 origin or faster evolutionary rate. As illustrated by the additional C_4-related genes upregulated in the C_4 plants from the Philippines, the rudimentary C_4 trait of *A. semialata* is likely to undergo similar secondary adaptations over evolutionary time.

Conclusions

In this study, the transcriptomes of individuals from the grass *Alloteropsis semialata* are analysed in a phylogenetic context to show that the changes in gene expression required for a physiological innovation can be spread over time. The relatively few changes required for the initial emergence of a metabolic pathway contrasted with the numerous modifications involved in the adaptation of this new pathway. Indeed, the emergence of a weak C_4 cycle in our study system was accompanied by the upregulation of three enzymes with a known C_4 function and 55 others proteins. The evolution of a stronger C_4 cycle then involved the upregulation of one other C_4 enzyme and 14 other proteins. However, adaptation of C_4 photosynthesis, illustrated here by population-specific expression of C_4-specific enzymes, continues when the plants are already in a C_4 state. The evolutionary modifications required to generate a rudimentary C_4 pathway can therefore be modest in species possessing C_4 enablers, but even a suboptimal C_4 pathway is important because it changes the environmental responses of the species. This creates an opportunity for natural selection to act on the standing variation, new mutations and, in some cases, laterally acquired genes, to assemble a trait of increasing complexity, allowing the colonization and gradual dominance in a larger spectrum of ecological conditions.

Data deposition

All raw DNA sequencing data (Illumina reads) and transcriptome assemblies generated as part of this study have been deposited with NCBI under Bioproject PRJNA401220.

Supplementary Data

Supplementary data are available at *JXB* online.

Table S1: List of enzymes considered as core C_4 enzymes.

Table S2: Information for populations sampled in triplicates.

Table S3: RNA-Seq data and mapping statistics for ten populations with triplicates.

Table S4: Pairwise differential expression test results for all genes.

Table S5: Leaf abundance, annotation, and summary of significance for all genes.

Table S6: Summary of differentially expressed genes referred to in Fig. 1.

Table S7: Summary of differentially expressed genes referred to in Fig. S1A.

Table S8: Summary of differentially expressed genes referred to in Fig. S1B.

Figure S1: Phylogenetic patterns of changes in gene expression in (A) *Alloteropsis angusta*, and (B) *Alloteropsis cimicina*.

Acknowledgements

This paper is dedicated to the memory of Mary Ann Cajano, from the University of the Philippines at Los Banos, who helped with the identification of plant specimens. The authors thank John Thompson who helped with plant collection. This work was funded by the Royal Society University Research Fellowship (grant number URF120119) and the Royal Society Research Grant (grant number RG130448) to PAC. LTD is funded by a NERC grant (grant number NE/M00208X/1), and JKO and MRL are funded by an ERC grant (grant number ERC-2014-STG-638333).

Author contributions

LTD, JJMV, AB, CPO, and PAC designed the research; LTD, MRL, JD, PS, CA, FN, JKO, AM, IMA, CJK, LAD, FK, JT, GB, WPQ, CPO, and PAC identified and collected plant material; LTD and JJMV generated and analysed the transcriptome data, with the help of AB and PAC; LTD, JJMV, and PAC wrote the paper with the help of all co-authors.

References

Aubry S, Brown NJ, Hibberd JM. 2011. The role of proteins in C_3 plants prior to their recruitment into the C_4 pathway. Journal of Experimental Botany **62**, 3049-3059.

Bläsing OE, Westhoff P, Svensson P. 2000. Evolution of C_4 phosphoenolpyruvate carboxylase in *Flaveria*, a conserved Serine residue in the carboxyl-terminal part of the enzyme is a major determinant for C_4-specific characteristics. Journal of Biological Chemistry **275**, 27917-27923.

Blount ZD, Barrick JE, Davidson CJ, Lenski RE. 2012. Genomic analysis of a key innovation in an experimental *Escherichia coli* population. Nature **489**, 513-518.

Borba AR, Serra TS, Górska A, Gouveia P, Cordeiro AM, Reyna-Llorens I, Kneřová J, Barros PM, Abreu IA, Oliveira MM *et al.* 2018. Synergistic binding of bHLH transcription factors to the promoter of the maize NADP-ME gene used in C_4 photosynthesis is based on an ancient code found in the ancestral C_3 state. Molecular Biology and Evolution **35,** 1690-1705.

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

Bräutigam A, Gowik U. 2016. Photorespiration connects C_3 and C_4 photosynthesis. Journal of Experimental Botany **67**, 2953-2962.

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.

Cao C, Xu J, Zheng G, Zhu X-G. 2016. Evidence for the role of transposons in the recruitment of *cis*-regulatory motifs during the evolution of C_4 photosynthesis. BMC Genomics **17**, 201**.**

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Molecular Biology and Evolution **17**, 540–552.

Chen T, Zhu XG, Lin Y. 2014. Major alterations in transcript profiles between C_3–C_4 and C_4 photosynthesis of an amphibious species *Eleocharis baldwinii*. Plant Molecular Biology, **86**, 93-110.

Christin PA, Boxall SF, Gregory R, Edwards EJ, Hartwell J, Osborne CP. 2013a. Parallel recruitment of multiple genes into C_4 photosynthesis. Genome Biology and Evolution **5**, 2174-

2187.

Christin PA, Osborne CP, Sage RF, Arakaki M, Edwards EJ. 2011. C_4 eudicots are not younger than C_4 monocots. Journal of Experimental Botany **62**, 3171-3181.

Christin PA, Osborne CP, Chatelet DS, Columbus JT, Besnard G, Hodkinson TR, Garrison LM, Vorontsova MS, Edwards EJ. 2013b. Anatomical enablers and the evolution of C_4 photosynthesis in grasses. Proceedings of the National Academy of Sciences, USA **110**, 1381–1386.

Christin PA, Spriggs E, Osborne CP, Strömberg CAE, Salamin N, Edwards EJ. 2014. Molecular dating, evolutionary rates, and the age of the grasses. Systematic Biology **63**, 153-165.

Christin PA, Osborne CP. 2014. The evolutionary ecology of C_4 plants. New Phytologist **204**, 765-781.

Christin PA, Arakaki M, Osborne CP, Edwards EJ. 2015. Genetic enablers underlying the clustered evolutionary origins of C_4 photosynthesis in angiosperms. Molecular Biology and Evolution **32**, 846-858.

Csárdi G, Franks A, Choi DS, Airoldi EM, Drummond DA. 2015. Accounting for experimental noise reveals that mRNA levels, anplified by post-transcriptional processes, largely determine steady-state protein levels in yeast. Plos Genetics **11**, e1005206.

Darwin C. 1859. On the origin of species by means of natural selection. Murray, London.

Dawkins R. 1986. The blind watchmaker. Norton, New York.

Ding Z, Weissmann S, Wang M, Du B, Huang L, Wang L, Tu X, Zhong S, Myers C, Brutnell TP *et al.* 2016. Identification of photosynthesis-associated C_4 candidate genes through comparative leaf gradient transcriptome in multiple lineages of C_3 and C_4 species. Plos One **10,** e0140629.

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

Dunning LT, Hipperson H, Baker WJ, Butlin RK, Devaux C, Hutton I, Igea J, Papadopulos AS, Quan X, Smadja CM, Turnbull CG, Savolainen V. 2016. Ecological speciation in sympatric palms: 1. Gene expression, selection and pleiotropy. Journal of Evolutionary Biology **29**, 1472-1487.

Dunning LT, Lundgren MR, Moreno-Villena JJ, Namaganda M, Edwards EJ, Nosil P, Osborne

CP, Christin PA. 2017. Introgression and repeated co-option facilitated the recurrent emergence of C_4 photosynthesis among close relatives. Evolution **71**, 1541-1555.

Dunning LT, Olofsson JK, Parisod C, Choudhury RR, Moreno-Villena JJ, Yang Y, Dionora J, Quick WP, Park M, Bennetzen JL *et al.* In press. Lateral transfers of large DNA fragments spread functional genes among grasses. Proceedings of the National Academy of Sciences USA doi:10.1073/pnas.1810031116

Ebersberger I, Strauss S, von Haeseler A. 2009. HaMStR: Profile hidden markov model based search for orthologs in ESTs. BMC Evolutionary Biology **9**, 157.

Emms DM, Covshoff S, Hibberd JM, Kelly S. 2016. Independent and parallel evolution of new genes by gene duplication in two origins of C_4 photosynthesis provides new insight into the mechanism of phloem loading in C_4 species. Molecular Biology and Evolution, **33,** 1796-1806.

Fracasso A, Trindade LM, Amaducci S. 2016. Drought stress tolerance strategies revealed by RNA-Seq in two sorghum genotypes with contrasting WUE. BMC Plant Biology **16**, 115.

Frean ML, Barrett DR, Ariovich D, Wolfson M, Cresswell CF. 1983. Intraspecific variability in *Alloteropsis semialata* (R. Br.) Hitchc. Bothalia **14**, 901-903.

Furumoto T, Yamaguchi T, Ohshima-Ichie Y, Nakamura M, Tsuchida-Iwata Y, Shimamura M, Ohnishi J, Hata S, Gowik U, Westhoff P, Bräutigam A, Weber APM, Izui K. 2011. A plastidial sodium-dependent pyruvate transporter. Nature **476**, 472-475.

Gowik U, Brautigam A, Weber KL, Weber APM, Westhoff P. 2011. Evolution of C_4 photosynthesis in the genus *Flaveria*: How many genes and which genes does it take to make C_4? The 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. Nature Biotechnology **29**, 644-652.

Grison MS, Brocard L, Fouillen L, Nicolas W, Wewer V, Dörmann P, Nacir H, Benitez-Alfonso Y, Claverol S, Germain V *et al.* 2015. Specific membrane lipid composition is important for plasmodesmata function in *Arabidopsis*. The Plant Cell 27, 1228-1250.

Hatch MD. 1987. C_4 photosynthesis: a unique blend of modified biochemistry, anatomy and ultrastructure. Biochimica et Biophysica 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.

Heyduk K, Moreno-Villena JJ, Gilman I, Christin PA, Edwards EJ. In press. The genetics of convergent evolution: insights from plant photosynthesis. Nature Reviews Genetics

Hibberd JM, Covshoff S. 2010. The regulation of gene expression required for C_4 photosynthesis. Annual Review of Plant Biology **68**, 181-207.

Huang P, Brutnell TP. 2016. A synthesis of transcriptomic surveys to dissect the genetic basis of C4 photosynthesis. Current Opinion in Plant Biology **31**, 91-99.

Huang P, Studer AJ, Schnable JC, Kellogg EA, Brutnell TP. 2017. Cross species selection scans identify components of C_4 photosynthesis in the grasses. Journal of Experimental Botany **68**, 127-135.

Jacob F. 1977. Evolution and tinkering. Science **196**, 1161-1166.

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 Physiology **165**, 62-75.

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

Kaldenhoff R, Kai L, Uehlein N. 2014. Aquaporins and membrane diffusion of CO_2 in living organisms. Biochmica et Biophysica Acta **1840**, 1592-1595.

Koussounadis A, Langdon SP, Um IH, Harrison DJ, Smith VA. 2015. Relationship between differentially expressed mRNA and mRNA-protein correlations in a xenograft model system. Scientific Reports **5**, 10775.

Ku MSB, Monson RK, Littlejohn RO, Nakamoto H, Fisher DB, Edwards GE. 1983. Photosynthetic characteristics of C_3-C_4 intermediate *Flaveria* species I. Leaf anatomy, photosynthetic responses to O_2 and CO_2, and activities of key enzymes in the C_3 and C_4 pathways. Plant Physiology **71**, 944-948.

Ku MSB, Wu J, Dai Z, Scott RA, Chu C, Edwards GE. 1991. Photosynthetic and photorespiratory characteristics of *Flaveria* species. Plant Physiology **96**, 518-528.

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. The Plant Cell **26**, 3243-3260.

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

Lauterbach M, Schmidt H, Billakurthi K, Hankeln T, Westhoff P, Gowik U, Kadereit G. 2017. De novo transcriptome assembly and comparison of C_3, C_3-C_4, and C_4 species of tribe Salsoleae (Chenopodiaceae). Frontiers in Plant Science **8,** 1939.

Lenski RE, Ofria C, Pennock RT, Adami C. 2003. The evolutionary origin of complex features. Nature **423**, 139-144.

Li Y, Ma X, Zhao J, Xu J, Shi J, Zhu XG, Zhao Y, Zhang H. 2015. Developmental genetic mechanisms of C_4 syndrome based on transcriptome analysis of C_3 cotyledons and C_4 assimilating shoots in *Haloxylon ammodendron*. Plos One **10**, e0117175.

Lundgren MR, Osborne CP, Christin PA. 2014. Deconstructing Kranz anatomy to understand C_4 evolution. Journal of Experimental Botany **65**, 3357-3369.

Lundgren MR, Besnard G, Ripley BS, Lehmann CER, Chatelet DS, Kynast RG, Namaganda M, Vorontsova MS, Hall RC, Elia J *et al*. 2015. Photosynthetic innovation broadens the niche within a single species. Ecology Letters **18**, 1021-1029.

Lundgren MR, Christin PA, Gonzalez Escobar E, Ripley BS, Besnard G, Long CM, Hattersley PW, Ellis RP, Leegood RC, Osborne CP. 2016. Evolutionary implications of C_3-C_4 intermediates in the grass *Alloteropsis semialata*. Plant, Cell and Environment **39**, 1874-1885

Lundgren MR, Christin PA. 2017. Despite phylogenetic effects, C_3–C_4 lineages bridge the ecological gap to C_4 photosynthesis. Journal of Experimental Botany **68**, 241-254.

Lundgren MR, Dunning LT, Olofsson JK, Moreno-Villena JJ, Bouvier JW, Sage TL, Khoshravesh R, Sultmanis S, Stata M, Ripley BS *et al.* 2019. C_4 anatomy can evolve via a single developmental change. Ecology Letters **22**, 302-312.

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.

Meléndez-Hevia E, Waddell TG, Cascante M. 1996. The puzzle of the Krebs citric acid cycle: assembling the pieces of chemically feasible reactions, and opportunism in the design of metabolic pathways during evolution. Journal of Molecular Evolution **43**, 293-303.

Min XJ, Butler G, Storms R, Tsang A. 2005. OrfPredictor: predicting protein-coding regions in EST-derived sequences. Nucleic Acids Research **33**, W677-W680.

Monson RK, Moore B, Ku MSB, Edwards GE. 1986. Co-function of C_3- and C_4-photosynthetic pathways in C_3, C_4 and C_3-C_4 intermediate *Flaveria* species. Planta **168**, 493-502.

Monson RK, Moore BD. 1989. On the significance of C_3-C_4 intermediate photosynthesis to the evolution of C_4 photosynthesis. Plant, Cell and Environment **12**, 689-699.

Monson RK, Teeri JA, Ku MSB, Gurevitch J, Mets LJ, Dudley S. 1988. Carbon-isotope discrimination by leaves of *Flaveria* species exhibiting different amounts of C_3- and C_4-cycle co-function. Planta **174**, 145-151.

Moreno-Villena JJ, Dunning LT, Osborne CP, Christin PA. 2018. Highly expressed genes are preferentially co-opted for C_4 photosynthesis. Molecular Biology and Evolution **35**, 94-106.

Niklaus M, Kelly S. 2019. The molecular evolution of C_4 photosynthesis: opportunities for understanding and improving the world’s most productive plants. Journal of Experimental Botany **70**, 795-804.

Notredame C, Higgins DG, Heringa J. 2000. T-coffee: a novel method for fast and accurate multiple sequence alignment1. Journal of Molecular Biology **302**, 205-217.

Olofsson JK, Bianconi M, Besnard G, Dunning LT, Lundgren MR, Holota H, Vorontsova MS, Hidalgo O, Leitch IJ, Nosil P, Osborne CP, Christin PA. 2016. Genome biogeography reveals the intraspecific spread of adaptive mutations for a complex trait. Molecular Ecology **25**, 6107-6123.

Rao X, Lu N, Li G, Nakashima J, Tang Y, Dixon RA. 2016. Comparative cell-specific transcriptomics reveals differentiation of C_4 photosynthesis pathways in switchgrass and other C_4 lineages. Journal of Experimental Botany 67, 1649-1662.

Reeves G, Singh P, Rossberg TA, Sogbohossou D, Schranz ME, Hibberd JM. 2018. Natural variation within a species for traits underpinning C_4 photosynthesis. Plant Physiology **177**, 504-512.

Reyna-Llorens I, Hibberd JM. 2017. Recruitment of pre-existing networks during the evolution of C_4 photosynthesis. Philosophical Transactions of the Royal Society, Series B **372**, 20160386.

Reyna-Llorens I, Burgess SJ, Reeves G, Singh P, Stevenson SR, Williams BP, Stanley S, Hibberd JM. 2018. Ancient duons may underpin spatial patterning of gene expression in C_4 leaves. Proceedings of the National Academy of Sciences USA **115**, 1931-1936.

Roberts A, Pachter L. 2013. Streaming fragment assignment for real-time analysis of sequencing experiments. Nature Methods **10**, 71-73.

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

Sage RF. 2004. The evolution of C_4 photosynthesis. New Phytologist **161**, 341-370.

Sage RF, Christin P-A, Edwards EJ. 2011. The C_4 plant lineages of planet Earth. Journal of Experimental Botany **62**, 3155–3169.

Sage RF, Sage TL, Kocacinar F. 2012. Photorespiration and the evolution of C_4 photosynthesis. Annual Review of Plant Biology **63**, 19-47.

Sage RF, Monson RK, Ehleringer JR, Adachi S, Pearcy RW. 2018. Some like it hot: The physiological ecology of C_4 plant evolution. Oecologia **187**, 941-966.

Sage TL, Busch FA, Johnson DC, Friesen PC, Stinson CR, Stata M, Sultmanis S, Rahman BA, Rawsthorne S, Sage RF. 2013. Initial events during the evolution of C_4 photosynthesis in C_3 species of Flaveria. Plant Physiology **163**, 1266-1276.

Schlüter U, Weber AP. 2016. The road to C_4 photosynthesis: evolution of a complex trait via intermediary states. Plant and Cell Physiology **57**, 881-889.

Schwanhäusser B, Busse D, Dittmar G, Schuchhardt J, Wolf J, Chen W, Selbach M. 2011. Global quantification of mammalian gene expression control. Nature **473**, 337-342.

Shen C, Li D, He R, Fang Z, Xia Y, Gao J, Shen H, Cao M. 2014. Comparative transcriptome analysis of RNA-Seq data for cold-tolerant and cold-sensitive rice genotypes under cold stress. Journal of Plant Biology **57**, 337-348.

Sonnhammer ELL, Ostlund G. 2014. InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. Nucleic Acids Research **43**, D234–D239.

South PF, Walker BJ, Cavanagh AP, Rolland V, Badger M, Ort DR. 2017. Bile Acid Sodium

Symporter BASS6 Can Transport Glycolate and Is Involved in Photorespiratory Metabolism in *Arabidopsis thaliana*. The Plant Cell **29**, 808-823.

Ueno O, Sentoku N. 2006. Comparison of leaf structure and photosynthetic characteristics of C_3 and C_4 *Alloteropsis semialata* subspecies. Plant, Cell and Environment **29**, 257-268.

von Caemmerer S. 1992. Stable carbon isotope discrimination in C_3–C_4 intermediates. Plant, Cell and Environment **15**, 1063-1072.

Vopalensky P, Pergner J, Liegertova M, Benito-Gutierrez E, Arendt D, Kozmik Z. 2012. Molecular analysis of the amphioxus frontal eye unravels the evolutionary origin of the retina and pigment cells of the vertebrate eye. Proceedings of the National Academy of Sciences USA **109**, 15383-15388.

Wang P, Khoshravesh R, Karki S, Tapia R, Balahadia CP, Bandyopadhyay A, Quick WP, Furbank R, Sage TL, Langdale JA. 2017. Re-creation of a key step in the evolutionary switch from C_3 to C_4 leaf anatomy. Current Biology **27**, 3278-3287.

Wang Y, Brautigam A, Weber APM, Zhu XG. 2014. Three distinct biochemical subtypes of C_4 photosynthesis? A modelling analysis. Journal of Experimental Botany 65, 3567-3578.

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.

Werner GD, Cornwell WK, Sprent JI, Kattge J, Kiers ET. 2014. A single evolutionary innovation drives the deep evolution of symbiotic N_2-fixation in angiosperms. Nature Communications **5**, 4087.

Yin X, Struik PC. 2018. The energy budget in C_4 photosynthesis: insights from a cell-type-specific electron transport model. New Phytologist **218**, 986-998.

Table 1: List of genes differentially expressed in key comparisons within *Alloteropsis semialata* from C_3 to C_3+C_4, and C_3+C_4 to C_4 that have SwissProt annotations. SwissProt protein description and *Arabidopsis* ortholog information is based on top-hit blast matches. Mean rpkm is derived from the seven *A. semialata* populations used for differential expression analysis (full summary of results can be found in Table S6).

		Arabidopsis	**Mean rpkm**		
Gene	**SwissProt protein description**	**ortholog**	**C_3**	**C_3+C_4**	**C_4**
Genes upregulated in C_3+C_4 and C_4 A. semialata (branch E in Fig. 4)					
ASEM_AUS1_17510[a]	Phosphoenolpyruvate carboxykinase (PCK)	AT4G37870	2	1168	3017
ASEM_AUS1_08268[a]	Aspartate aminotransferase (ASP-AT)	AT5G11520	158	1843	1196
ASEM_AUS1_19029[a]	Phosphoenolpyruvate carboxylase (PEPC)	AT2G42600	95	828	1118
ASEM_AUS1_30031[a]	Fruit bromelain	AT1G06260	11	260	497
ASEM_AUS1_08709	Iron-sulfur cluster assembly protein 1	AT4G22220	67	394	473
ASEM_AUS1_11198	Bifunctional TENA2 protein	AT3G16990	10	43	80
ASEM_AUS1_19914	50S ribosomal protein L17	AT5G64650	1	78	58
ASEM_AUS1_02887[a]	Cysteine proteinase 1	AT2G32230	0	44	54
ASEM_AUS1_16281[a]	Probable carboxylesterase 15	AT5G06570	1	16	50
ASEM_AUS1_11666	Putative protease Do-like 14	AT5G27660	1	63	39
ASEM_AUS1_18766[a]	Nudix hydrolase 16	AT3G12600	4	24	38
ASEM_AUS1_21431[a]	DNA-binding protein MNB1B	AT4G35570	0	94	30
ASEM_AUS1_24040[a,b]	Putative phosphatidylglycerol/phosphatidylinositol transfer protein	AT3G11780	4	32	24
ASEM_AUS1_08934	Putative F-box protein	AT4G38870	0	18	23
ASEM_AUS1_44075	Indole-3-acetaldehyde oxidase	AT5G20960	0	28	22
ASEM_AUS1_24692	Dihydrolipoyllysine-residue acetyltransferase component 1 of pyruvate dehydrogenase complex	AT3G52200	0	13	20
ASEM_AUS1_38810	UDP-glycosyltransferase	AT1G05680	0	35	17
ASEM_AUS1_24427	Putative F-box protein	AT1G65770	0	19	16
ASEM_AUS1_43609[a]	Flavin-containing monooxygenase FMO GS-OX-like 9	AT5G07800	0	7	13
ASEM_AUS1_40960	Cysteine-rich receptor-like protein kinase 26	AT4G23240	1	18	13
ASEM_AUS1_16960[a]	Valine--tRNA ligase	AT1G14610	0	26	12
ASEM_AUS1_27461[b]	Aspartic proteinase nepenthesin-2	AT2G03200	0	2	12
ASEM_AUS1_15840	Tyrosine--tRNA ligase	AT2G33840	0	4	10
ASEM_AUS1_22664	Probable nucleolar protein 5-1	AT5G27120	0	19	8
ASEM_AUS1_39034	Putative protease Do-like 14	AT5G27660	0	11	7
ASEM_AUS1_21913	Protein NEN1	AT5G07710	0	5	6
ASEM_AUS1_01903	Disease resistance protein RPM	AT3G07040	0	7	2
Genes downregulated in C_3+C_4 and C_4 A. semialata (branch E in Fig. 4)					
ASEM_AUS1_21734	60S ribosomal protein L23a	AT3G55280	206	0	72
ASEM_AUS1_01414[a,b]	Acyl transferase 4	AT3G62160	150	18	17
ASEM_AUS1_31537	Pumilio homolog 23	AT1G72320	49	12	9
ASEM_AUS1_00061	40S ribosomal protein SA	AT3G04770	42	7	7
ASEM_AUS1_22162	Tubulin alpha-3 chain	AT4G14960	32	6	3
ASEM_AUS1_22449[a]	Callose synthase 3	AT5G13000	30	2	1
ASEM_AUS1_04268[a]	40S ribosomal protein S21	AT5G27700	20	0	0
ASEM_AUS1_06562[a,b]	PTI1-like tyrosine-protein kinase 3	AT3G59350	5	1	1
Genes upregulated in C_4 A. semialata (branch I in Fig.4)					
ASEM_AUS1_39556[a,b]	Pyruvate, phosphate dikinase 1 (PPDK)	AT4G15530	60	133	1149
ASEM_AUS1_24184[a]	Phosphatidylglycerol/phosphatidylinositol transfer protein	AT3G11780	0	1	104
ASEM_AUS1_29700	Protein SRG1	AT1G17020	2	1	86
ASEM_AUS1_16577[a]	Lactoylglutathione lyase	AT1G11840	0	0	46
ASEM_AUS1_06220	S-norcoclaurine synthase 1	AT1G17020	1	1	39
ASEM_AUS1_24241	DnaJ homolog subfamily A member 1	AT3G14200	1	1	33

ASEM_AUS1_44200[a]	Aquaporin TIP1-1	AT2G36830	0	0	17
ASEM_AUS1_13652	Transcription factor TGAL4	AT1G08320	0	0	7
ASEM_AUS1_00246	Nicotinamide adenine dinucleotide transporter 2	AT1G25380	0	0	2
Genes downregulated in C_4 A. semialata (branch I in Fig.4)					
ASEM_AUS1_43847[a,b]	Short-chain dehydrogenase TIC 32	AT4G23420	18	11	0

[a] Significant change in the same direction in *A. angusta*; [b] Significant change in the same direction in *A. cimicina*

Figure captions

Figure 1: Phylogenetic tree inferred from multiple nuclear markers.

(A) This phylogeny was inferred under maximum likelihood using transcriptome-wide markers. Scale indicates number of nucleotide substitutions per site, and bootstrap support values are indicated near nodes. AANG = A. angusta. For *A. semialata*, population names indicate the country of origin; AUS = Australia, BUR = Burkina Faso, CMR = Cameroon, MAD = Madagascar, PHI = Philippines, RSA = South Africa, TAN = Tanzania, SRI = Sri Lanka, TPE = Chinese Taipei, ZAM = Zambia, ZIM = Zimbabwe. Populations sampled with biological replicates and used for differential expression analysis are indicated by the large circles and bold population names. Nuclear clades from Olofsson *et al.* (2016) are indicated. Branch colours indicate the ancestral photosynthetic types, based on the transcriptomes and leaf anatomy detailed investigations of Dunning *et al.* (2017). The hashed green at the base of *A. semialata* indicates uncertainty between C_3 and C_3+C_4 states. (B) Distribution of *A. semialata photosynthetic* types and sampling locations, with color codes as in panel A. Shadings indicate the approximate ranges of the three photosynthetic types of *A. semialata*, based on Lundgren *et al.* (2016).

Figure 2: Expression profile similarity across all samples.

Expression profiles are clustered in multidimensional scaling (MDS) plots using (A) all samples (B) only *A. semialata* samples. Species and nuclear clades from Olofsson *et al.* (2016) are delimited and population names are as in Fig. 1.

Figure 3: Number of differentially expressed genes among pairs of populations.

The heatmap shows the number of significantly differentially expressed genes detected for each pair of populations. The phylogenetic relationships among populations are indicated on the side, using an ultrametric version of the tree presented in Fig. 1.

Figure 4: Phylogenetic patterns of changes in gene expression.

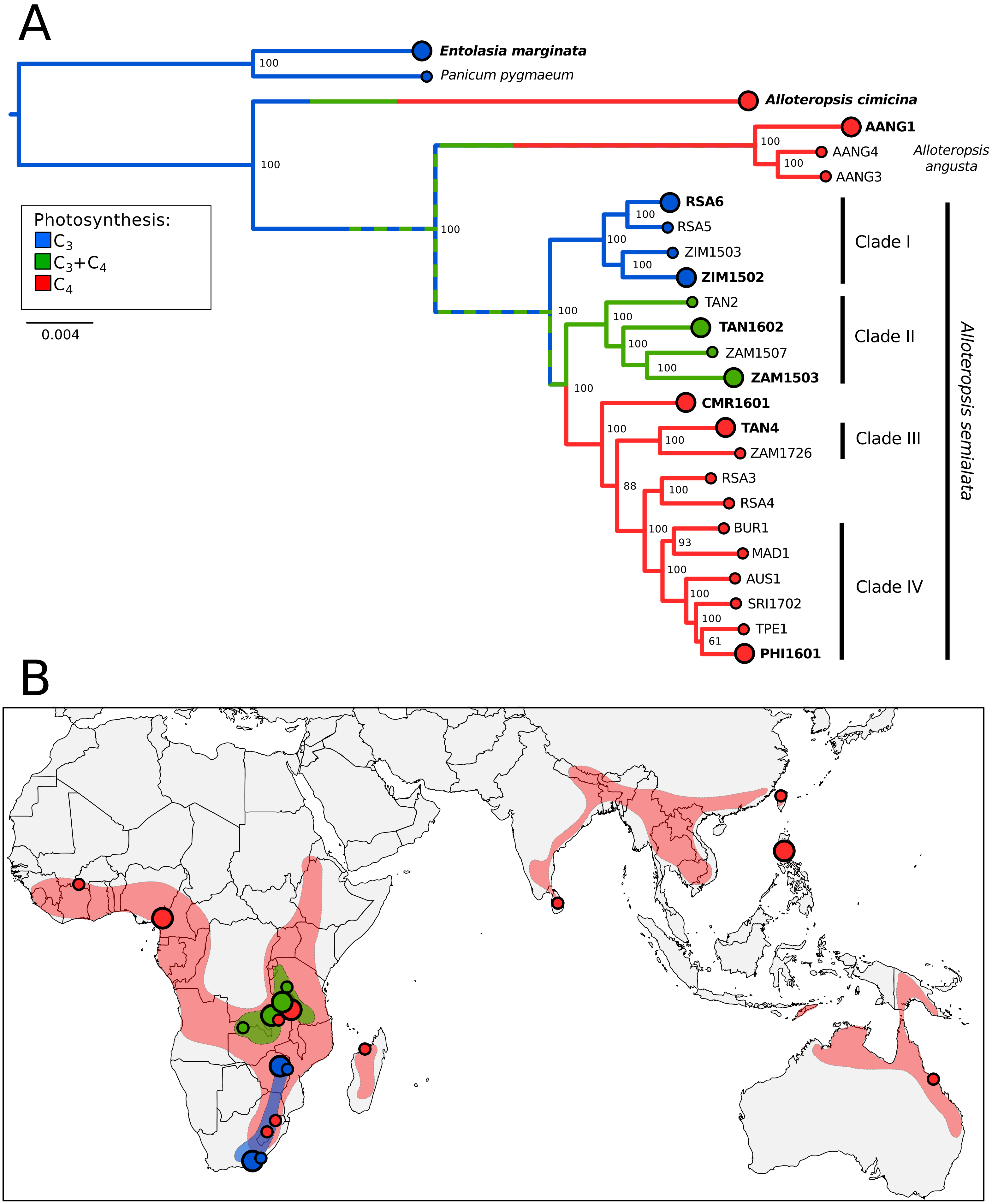
The maximum-likelihood phylogeny from Fig. 1 is shown unrooted after pruning the populations not used for expression analyses. For each branch, the number of differentially expressed genes is indicated, with numbers next to arrows indicating those that are consistently up- or down-regulated as you move along the tree from the outgroup *Entolasia marginata*. Each population has three biological replicates, and colours indicate the photosynthetic type (blue = C_3; green = C_3+C_4; red = C_4). Scale indicates number of nucleotide substitutions per site, with truncated branches highlighted by two bars. The two greyed out C_4 congeners were excluded from these analyses, and results that involve them can be found in Fig. S3.

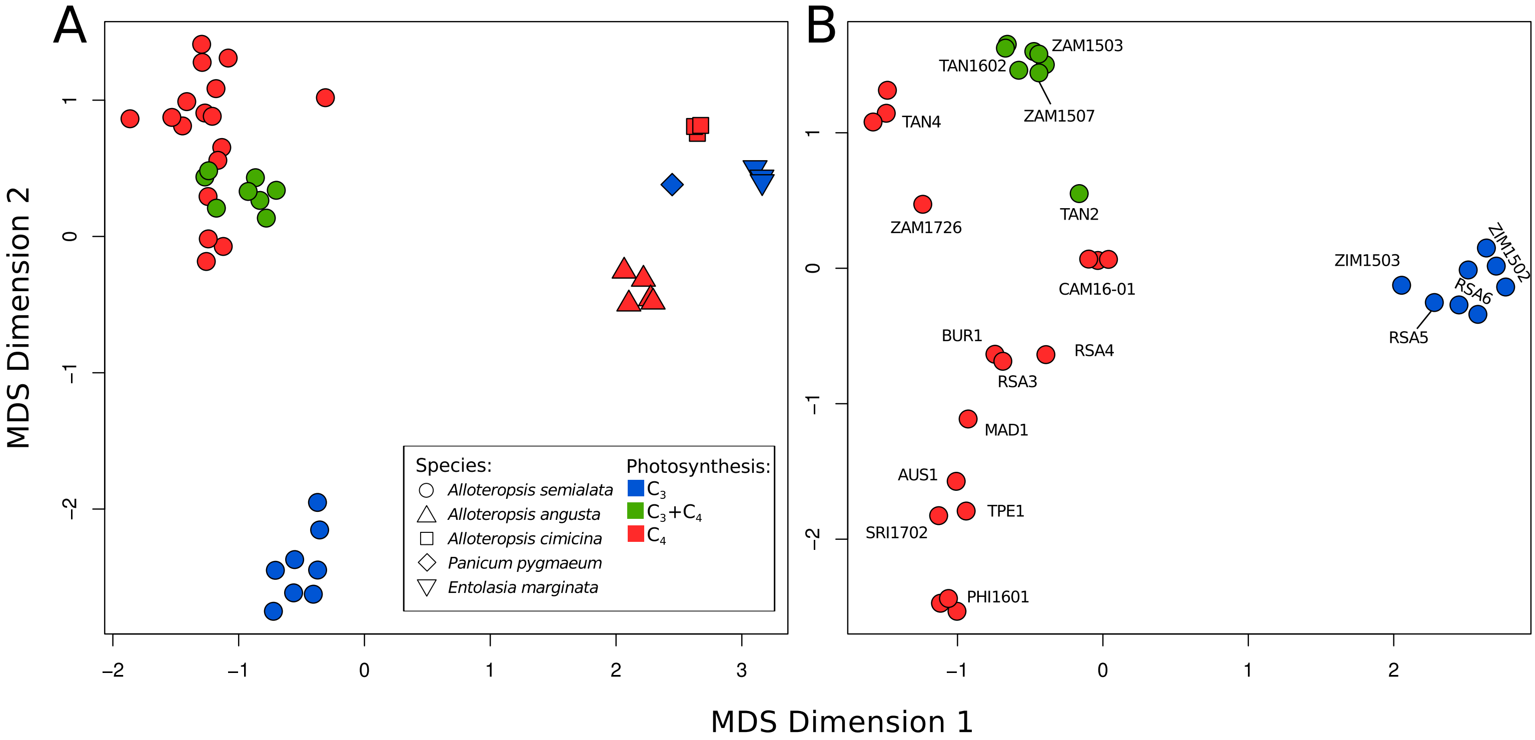
Figure 5: Expression levels of exemplar genes across accessions.

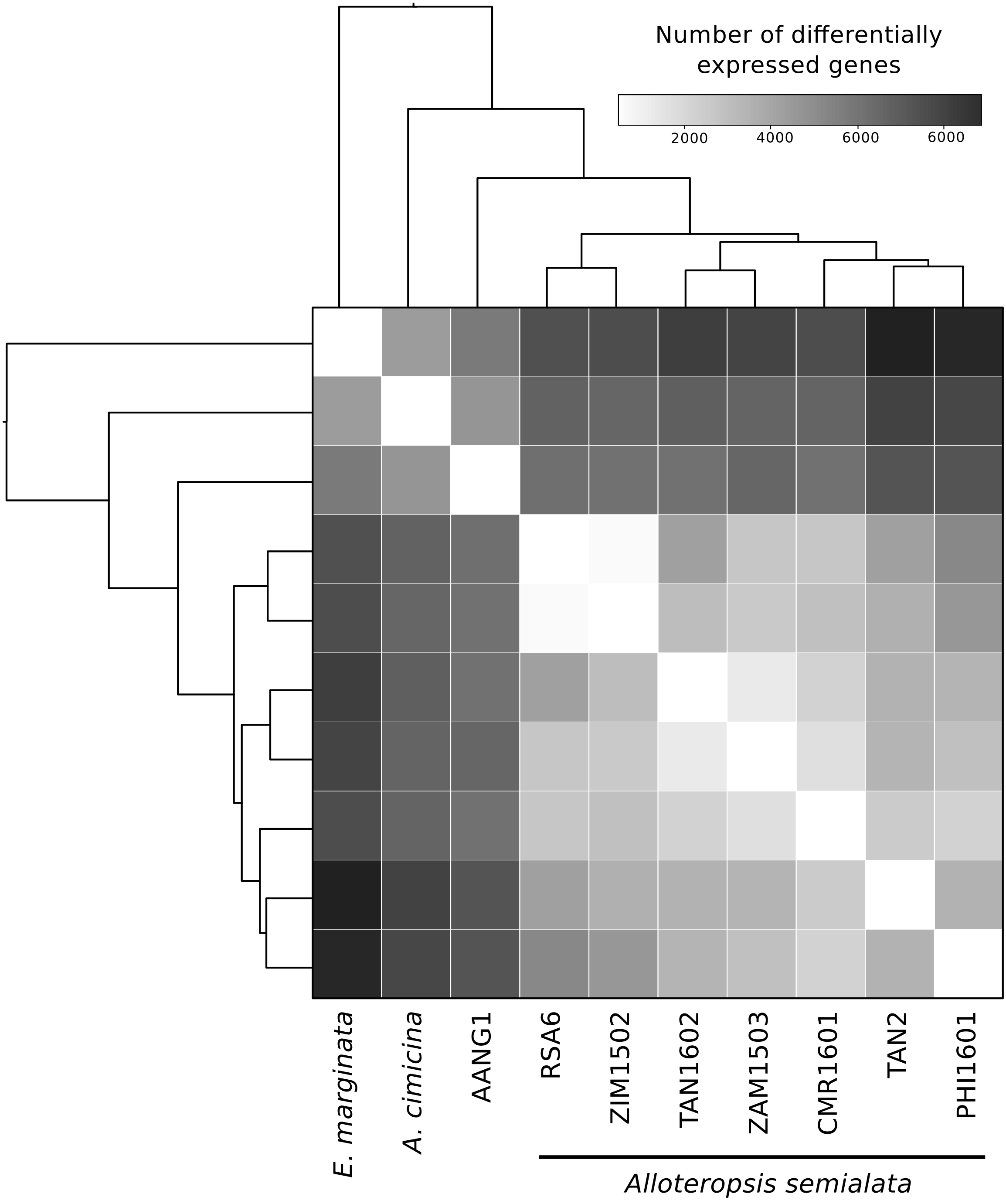
Expression levels in reads per kilobase of transcript per million mapped reads (rpkm) are shown for four example genes. Standard deviation for populations with biological replicates is indicated. Colours indicate the photosynthetic types; blue = C_3; green = C_3+C_4; red = C_4.

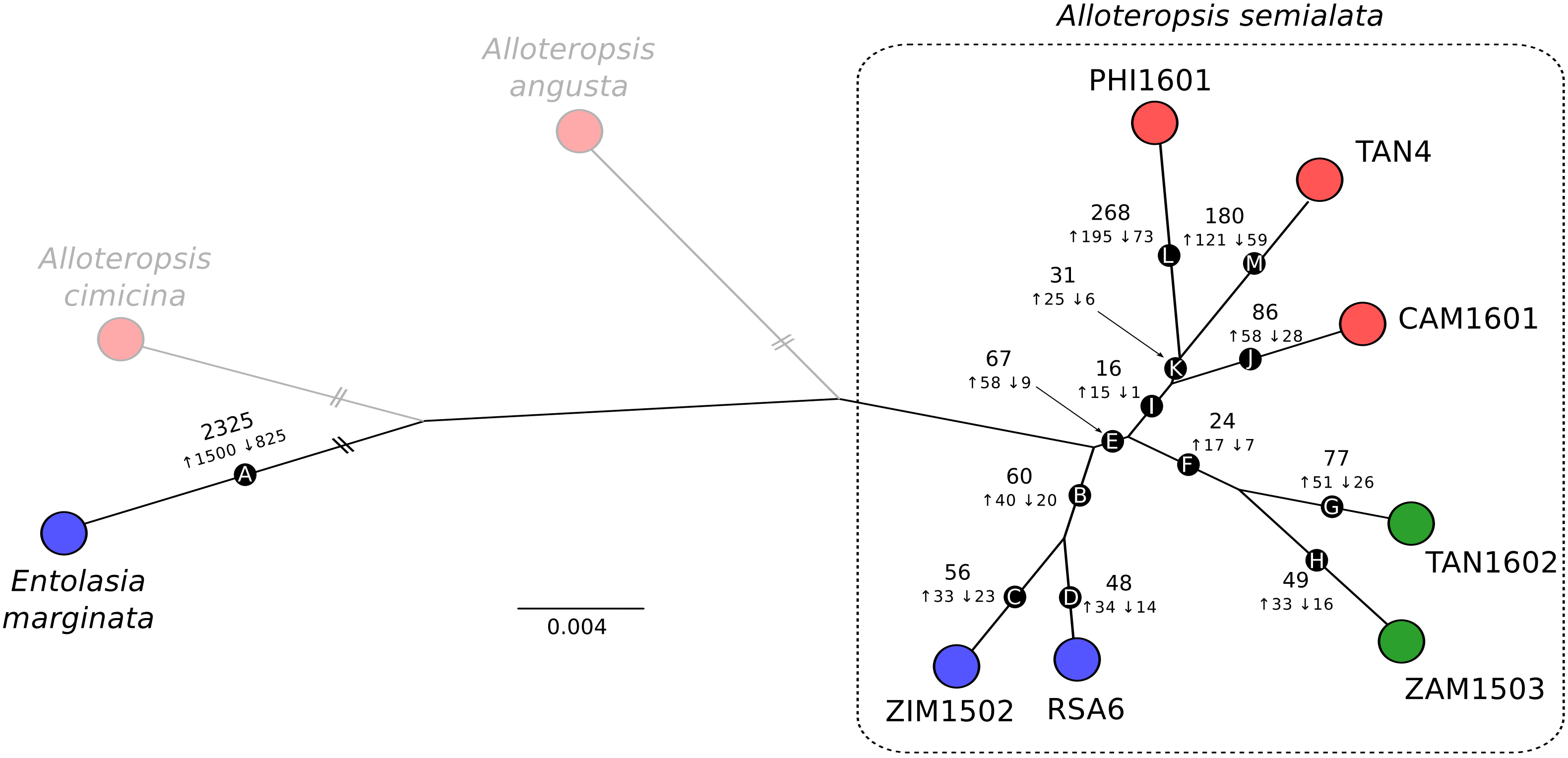
Figure 6: Putative C_4 pathway in *Alloteropsis semialata*

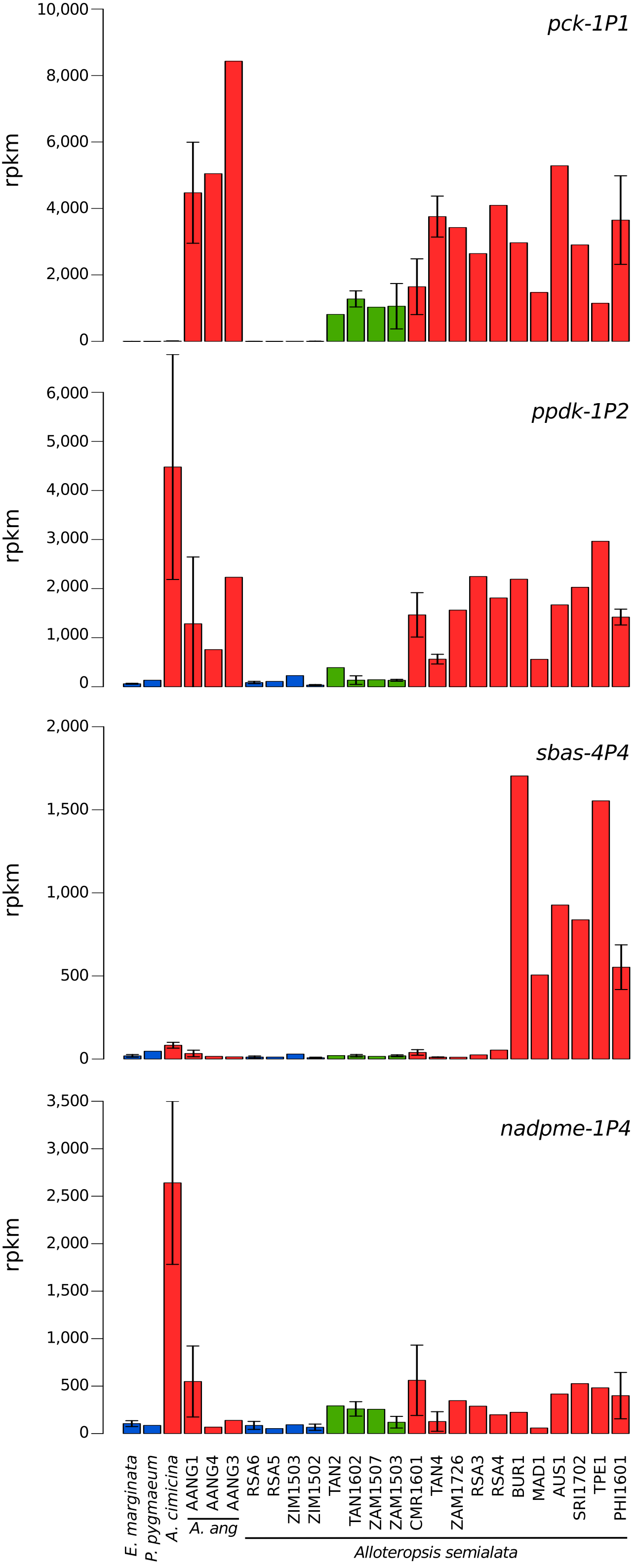
A C_4 cycle is suggested for *A. semialata* based on the transcript abundance of C_4-related genes, and the literature (Frean *et al.*, 1983; Ueno & Sentoku, 2006). Pathway components are coloured per the differential expression analysis, with those in black being putatively sufficiently abundant in C_3 ancestors, parts of the pathway in green upregulated during the transition to C_3+C_4, and parts in red upregulated during the transition from C_3+C_4 to C_4. ALA-AT = alanine aminotransferase, ASP-AT = aspartate aminotransferase, CA = carbonic anhydrase, NADP-MDH = NADP malate dehydrogenase, NAD(P)-ME = NAD(P) malic enzyme, PCK = phosphoenolpyruvate carboxykinase, PEPC = phosphoenolpyruvate carboxylase, PEPP = phosphoenolpyruvate phosphatase, PPDK = pyruvate orthophosphate dikinase, PCR = photosynthetic carbon reduction (Calvin-Benson cycle).

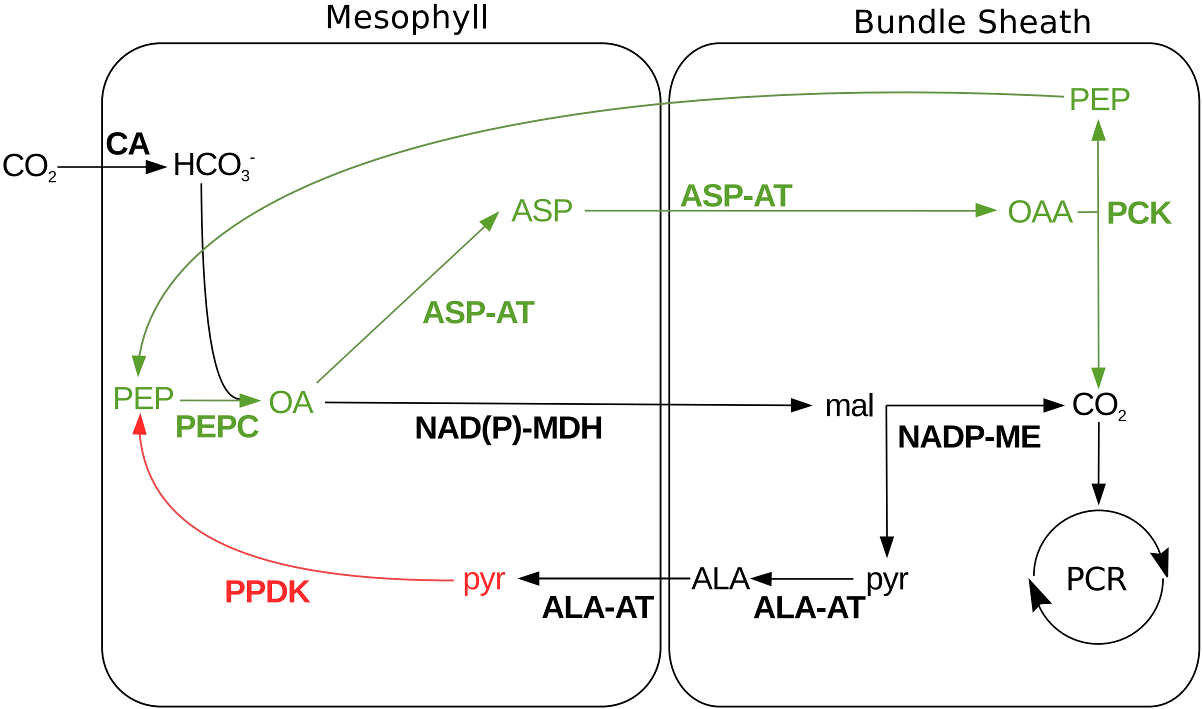

A
Entolasia marginata
Panicum pygmaeum
Alloteropsis cimicina
AANG1
AANG4
AANG3
Alloteropsis angusta
RSA6
RSA5
ZIM1503
ZIM1502
TAN2
TAN1602
ZAM1507
ZAM1503
CMR1601
TAN4
ZAM1726
RSA3
RSA4
BUR1
MAD1
AUS1
SRI1702
TPE1
PHI1601
Clade I
Clade II
Clade III
Clade IV
Alloteropsis semialata
100
100
100
100
100
100
100
100
100
100
100
100
100
100
100
100
88
100
100
93
100
100
100
61
Photosynthesis:
🟦 C_3
🟩 C_3+C_4
🟥 C_4
0.004
B


A
B
MDS Dimension 2
MDS Dimension 1
−2
−1
0
1
2
3
Species:
○ Alloteropsis semialata
△ Alloteropsis angusta
□ Alloteropsis cimicina
◇ Panicum pygmaeum
▽ Entolasia marginata
Photosynthesis:
🟦 C_3
🟩 C_3+C_4
🟥 C_4
ZAM1503
TAN1602
ZAM1507
TAN4
ZAM1726
TAN2
CAM16-01
ZIM1503
ZIM1502
RSA6
RSA5
BUR1
RSA4
RSA3
MAD1
AUS1
TPE1
SRI1702
PHI1601


Number of differentially
expressed genes
2000
4000
6000
6000
E. marginata
A. cimicina
AANG1
RSA6
ZIM1502
TAN1602
ZAM1503
CMR1601
TAN2
PHI1601
Alloteropsis semialata


Alloteropsis semialata
Alloteropsis angusta
Alloteropsis cimicina
Entolasia marginata
PHI1601
TAN4
CAM1601
TAN1602
ZAM1503
ZIM1502
RSA6
2325
↑1500 ↓825
A
268
↑195 ↓73
L
180
↑121 ↓59
M
31
↑25 ↓6
K
86
↑58 ↓28
J
67
↑58 ↓9
E
16
↑15 ↓1
I
24
↑17 ↓7
F
77
↑51 ↓26
G
60
↑40 ↓20
B
56
↑33 ↓23
C
48
↑34 ↓14
D
49
↑33 ↓16
H
0.004

pck-1P1
ppdk-1P2
sbas-4P4
nadpme-1P4
rpkm
10,000
8,000
6,000
4,000
2,000
0
6,000
5,000
4,000
3,000
2,000
1,000
0
2,000
1,500
1,000
500
0
3,500
3,000
2,500
2,000
1,500
1,000
500
0
E. marginata
P. pygmaeum
A. cimicina
AANG1
AANG4
AANG3
A. ang
RSA6
RSA5
ZIM1503
ZIM1502
TAN2
TAN1602
ZAM1507
ZAM1503
CMR1601
TAN4
ZAM1726
RSA3
RSA4
BUR1
MAD1
AUS1
SRI1702
TPE1
PHI1601
Alloteropsis semialata


Mesophyll
Bundle Sheath
PEP
CA
CO2
HCO3-
ASP
ASP-AT
OAA
PCK
ASP-AT
PEP
OA
mal
CO2
PEPC
NAD(P)-MDH
NADP-ME
PCR
pyr
ALA
pyr
PPDK
ALA-AT
ALA-AT