Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

The Coordination and Jumps along C4 Photosynthesis Evolution in the Genus Flaveria

Amy Lyu
CAS Center for Excellence in Molecular Plant Sciences
Udo Gowik
Heinrich Heine University Düsseldorf
Steve Kelly
University of Oxford
Sarah Covshoff
University of Cambridge
Harmony Clayton
University of Western Australia
Julian Hibberd
University of Cambridge
Rowan Sage
University of Toronto
Martha Ludwig
University of Western Australia
Gane Gane
BGI-Shenzhen, Beishan Industrial Zone, Yantian District, Shenzhen 518083, China
Peter Westhoff
Heinrich Heine University Düsseldorf
Xin-Guang Zhu (zhuxg@cemps.ac.cn)
CAS Center for Excellence in Molecular Plant Sciences

Research Article

Keywords: C4, photosynthesis, evolution, coordination, jump, Flaveria

Posted Date: January 11th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-140156/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Version of Record: A version of this preprint was published at Scientific Reports on August 2nd, 2021. See the published version at https://doi.org/10.1038/s41598-021-93381-8.

Title: The Coordination and Jumps along C_4 Photosynthesis Evolution in the Genus *Flaveria*

Author information

Ming-Ju Amy Lyu： National Key Laboratory of Plant Molecular Genetics, CAS Center for Excellence in Molecular Plant Sciences, Institute of Plant Physiology and Ecology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China, lvmj@cemps.ac.cn.

Udo Gowik: Institute of Plant Molecular and Developmental Biology, Heinrich-Heine-University, Dusseldorf, Germany, gowik@uni-duesseldorf.de.

Steve Kelly: Department of Plant Sciences, University of Oxford, Oxford, United Kingdom, steven.kelly@plants.ox.ac.uk.

Sarah Covshoff: Department of Plant Sciences, University of Cambridge, Cambridge, United Kingdom, sarahcovshoff@gmail.com.

Harmony Clayton: School of Molecular Sciences, University of Western Australia, Crawley, WA, Australia, 20152857@student.uwa.edu.au.

Julian M. Hibberd: Department of Plant Sciences, University of Cambridge, Cambridge, United Kingdom, jmh65@cam.ac.uk.

Rowan F. Sage: Department of Ecology and Evolutionary Biology, University of Toronto, Toronto, Canada, r.sage@utoronto.ca.

Martha Ludwig: School of Molecular Sciences, University of Western Australia, Crawley, WA, Australia, martha.ludwig@uwa.edu.au.

Gane Ka-Shu Wong: BGI-Shenzhen, Beishan Industrial Zone, Yantian District, Shenzhen 518083, China; Department of Biological Sciences, University of Alberta, Edmonton AB, T6G 2E9, Canada; Department of Medicine, University of Alberta, Edmonton AB, T6G 2E1, Canada, gane@ualberta.ca.

Peter Westhoff: Institute of Plant Molecular and Developmental Biology, Heinrich-Heine-University, Dusseldorf, Germany, west@uni-duesseldorf.de.

Xin-Guang Zhu*: National Key Laboratory of Plant Molecular Genetics, CAS Center for Excellence in Molecular Plant Sciences, Institute of Plant Physiology and Ecology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China, zhuxg@cemps.ac.cn

Author for Correspondence

Xin-Guang Zhu

National Key Laboratory for Molecular Plant Sciences, Institute of Plant Physiology and Ecology, Chinese Academy of Sciences, Shanghai, China, 300 Fenglin Road, Xuhui Area, Shanghai 200031, China

Email: zhuxg@sippe.ac.cn

Fax: 86-21-54920451

Tel: 86-21-54920486

The Coordination and Jumps along C_4 Photosynthesis Evolution in the Genus *Flaveria*

Ming-Ju Amy Lyu[1], Udo Gowik[2], Steve Kelly[3], Sarah Covshoff[4], Harmony Clayton[5], Julian M. Hibberd[4], Rowan F. Sage[6], Martha Ludwig[5], Gane Ka-Shu Wong[7,8,,9], Peter Westhoff[2], Xin-Guang Zhu[1§]

1. National Key Laboratory of Plant Molecular Genetics, CAS Center for Excellence in Molecular Plant Sciences, Institute of Plant Physiology and Ecology, Chinese Academy of Sciences, Shanghai, China

2. Institute of Plant Molecular and Developmental Biology, Heinrich-Heine-University, Dusseldorf, Germany

3. Department of Plant Sciences, University of Oxford, Oxford, United Kingdom

4. Department of Plant Sciences, University of Cambridge, Cambridge, United Kingdom

5. School of Molecular Sciences, University of Western Australia, Crawley, WA, Australia

6. Department of Ecology and Evolutionary Biology, University of Toronto, Toronto, Canada

7. BGI-Shenzhen, Beishan Industrial Zone, Yantian District, Shenzhen 518083, China

8. Department of Biological Sciences, University of Alberta, Edmonton AB, T6G 2E9, Canada

9. Department of Medicine, University of Alberta, Edmonton AB, T6G 2E1, Canad

[§]Corresponding author

Abstract

C_4 photosynthesis is a remarkable complex trait, elucidations of the evolutionary trajectory of C_4 photosynthesis from its ancestral C_3 pathway can help us better understand the generic principles of the evolution of complex trait and guide the engineering of C_3 crops for higher yields. Here, we used the genus *Flaveria* that contains C_3, C_3-C_4, C_4-like and C_4 species as a system to study the evolution of C_4 photosynthesis. We first mapped transcript abundance, protein sequence, and morphological features to the phylogenetic tree of the genus *Flaveria*, and calculated the evolutionary correlation of different features; we then predicted the relative changes of ancestral nodes of those features to illustrate the key stages during the evolution of C_4 photosynthesis. We found that gene expression and protein sequence showed consistent modification pattern along the phylogenetic tree. High correlation coefficients ranging from 0.46 to 0.9 among gene expression, protein sequence and morphology were observed, and the greatest modification of those different features consistently occurred at the transition between C_3-C_4 species and C_4-like species. Our results show highly coordinated changes in gene expression, protein sequence and morphological features, which support an obviously evolutionary jump during the evolution of C_4 metabolism.

Key words

C_4 photosynthesis; evolution; coordination; jump; *Flaveria*

Introduction

Elucidating the evolutionary and developmental processes of complex traits formation is a major focus of current biological and medical research. Most health related issues, including obesity and diabetes, as well as agricultural challenges, such as flowering time control, crop yield improvements, and disease resistance, are related to complex traits [15, 26, 32]. Currently, genome-wide association studies are applied in the study of complex traits. Putative genes or molecular markers are then evaluated by a reverse genetics approach to identify those influence the complex traits. C_4 photosynthesis is a complex trait that evolved from C_3 photosynthesis. When compared with C_3 plants, C_4 plants have higher water, nitrogen and light using efficiencies [50]. Interestingly, C_4 photosynthesis has evolved independently more than 66 times, representing a remarkable example of convergent evolution [37]. Accordingly, C_4 evolution is an ideal system for the investigation of the mechanisms of convergent evolution of complex traits.

C_4 photosynthesis contains a number of biochemical, cellular and anatomical modifications when compared with the ancestral C_3 photosynthesis [12, 36]. In C_3 photosynthesis, CO_2 is fixed by ribulose-1,5-bisphosphate carboxylase/oxygenase (Rubisco), whereas in dual-cell C_4 photosynthesis, CO_2 is initially fixed into a four-carbon organic acid in mesophyll cells (MCs) by phospho*enol*pyruvate carboxylase (PEPC) [14]. The resulting four-carbon organic acid then diffuses into the bundle-sheath cells (BSCs) [18], where CO_2 is released and fixed by Rubisco. Hence, C_4 photosynthesis requires extra enzymes in CO_2 fixation in addition to those already functioning in C_3 photosynthesis, including PEPC, NADP-dependent malic enzyme (NADP-ME), and pyruvate, orthophosphate dikinase (PPDK) [14]. In dual-cell C_4 photosynthesis, CO_2 is concentrated in BSCs that are surrounded by MCs, forming the

so-called Kranz anatomy [4, 13, 41]. Compared with C_3 leaf anatomy, Kranz anatomy requires a spatial rearrangement of MCs and BSCs, cell size adjustment for increased numbers of organelles, larger organelles and metabolite transfer between the two cell types, and a reduction in distance between leaf veins.

Much of the current knowledge regarding the evolution of C_4 photosynthesis was gained through comparative studies in terms of physiology and anatomy by using genera that have both C_3 and C_4 species, as well as species performing intermediate types of photosynthesis [36, 38]. Among these, the genus *Flaveria* has been promoted as a model for C_4 evolution studies [27], and the evolution of C_4-related morphological, anatomical and physiological features has been well studied in this genus over the last 40 years [3, 9, 24, 27]. The molecular evolution of several key C_4 enzymes have been reported in this genus [7, 8, 48], however, the molecular evolution of most C_4 related genes is largely unknown. Besides, the evolutionary relationship between the C_4 related genes and morphology features is not clear so far. In this study, we combined transcriptome data and published morphology data, together with the most recent phylogenetic tree of the genus *Flaveria* [22], to systematically investigate the key molecular events and evolutionary paths during the C_4 evolution. Our results revealed that many of the events related to C_4 photosynthesis occurred in a coordinated way and presented jumping changes along the process.

Results

Transcriptome assembly and quantification

RNA-Seq data of 31 samples of 16 *Flaveria* species were obtained from the public database

Sequence Read Achieve (SRA) of the National Center for Biotechnology Information (NCBI) (Table S1). The 16 species represented two C_3 species, seven C_3-C_4 intermediate species, three C_4-like species and four C_4 species [5, 20] (Table S1). On average, 42,132 contigs (from 30,968 to 48,969) were assembled with N50 ranging from 658 bp to 1208 bp among the 16 species (Table S2). The distribution of the contig length is similar in the 16 species with a peak at 360 bp (Fig. S1). Since *Flaveria* is a eudicot genus, we used *Arabidopsis thaliana* (Arabidopsis) as a reference to annotate *Flaveria* transcripts. On average, 58.91% of *Flaveria* contigs had orthologous genes in Arabidopsis.

To estimate the accuracy of the annotation, we used OrthoFinder [6] to predict the orthologous groups based on annotated genes of *Flaveria* species and then calculated the consistence between our gene annotation and orthologous groups. Specifically, for each orthologous group, we calculated the percentage of genes with the same annotation. For example, for the orthologous group of PPDK(AT4G15530), 29 transcripts from *Flaveria* and one gene (AT4G15530) with six transcripts from Arabidopsis were clustered in this orthologous group. All of the 29 transcripts from *Flaveria* were annotated as AT4G15530 (Fig. S2A). Therefore, the consistency of our annotation is 100% (29/29) for this orthologous group. Our result showed that 80% of the total 28,164 orthologous groups has a consistency higher than 90% (Fig. S2).

Transcript abundance was calculated as fragments per kilobase of transcript per million mapped reads (FPKM) (see Methods). The total transcriptome-level comparison revealed higher Pearson correlation in overall transcript abundance in leaf samples from the same species than that of different organs from the same species, regardless of sources (Fig. S3). Specifically, leaves from different developmental stages or from different labs are more closely correlated than leaf samples from different

species, or than mean values of pair-wise correlations across all 27 leaf samples (T-test, P<0.05) (Fig. S4). As a result, 13,081 Arabidopsis orthologs were detected in at least one of the 16 *Flaveria* species, and 12,215 were kept with the maximum FPKM in 16 species ≥ 1 FPKM.

Investigate the possibility of samples used in this study being hybrid

Considering that intermediacy of traits in the intermediate species may result from a possible hybridization between species with one parent being C_3 and another parent being C_4 or intermediate species [19], we investigated whether the intermediate species used in this study are naturally evolved intermediate species or a hybrid offspring. The hybrid offspring is characterized as expressing different alleles at one DNA site, therefore we calculated the percentage of DNA sites that expressed different alleles, which is termed as mixed site. The percentage of mixed site was then compared to the positive background generated by pair-wise mixing of RNA-Seq data of 16 *Flaveria* species. Our result showed that the known hybrid sample *F. pringlei** originated from *F. pringlei* ×*F. angustiflolia* in [22] showed significantly higher percentage of mixed site than background (Binomial test, P<0.001), whereas, other species showed significantly lower percentages of mixed sites than background (Binomial test, P<0.001) (Fig. 1). Thus, our data showed that all species used in this study are from natural evolution and can be used for evolutionary study.

We used the phylogenetic tree of the genus *Flaveria* [23] to illustrate the molecular evolution of C_4 photosynthesis. We numbered each node of the phylogenetic tree in an evolutionarily sequential order, namely, the more ancient a node, the lower its number. The number begins with N1 which refers to the common ancestor of all *Flaveria* species in the phylogenetic tree, N3 refers to the common ancestor on all intermediate and C_4 species, at which intermediate species first evolved. N7 is the common ancestor

of C_4-like and C_4 species in clade A, where a completed C_4 cycle evolved. The nodes of clade B were numbered sequentially following clade A.

The modified genes: genes showed difference in gene expression and protein sequence between C_3 species and C_4 species

We first identified the modified genes, which were defined as genes show difference in both gene expression and protein sequence between C_3 and C_4 species. We first calculated the differentially expressed (DE) genes between C_3 and C_4 species in the way of comparing three C_3 samples with 8 C_4 samples (see Method), which resulted in 896 DE genes ("BH" correlated P<0.05) (Additional file 3). Besides, 1222 DE genes were detected between C_3 and Type I C_3-C_4 species from clade A (*F. sonorensis*, *F. angustifolia*), and 318 DE genes between C_3 and all C_3-C_4 species from both clade A and clade B (Additional file 3). 117 genes were overlapped among the three datasets, suggesting the large variance among different C_3-C_4 species. Therefore, we used the 896 DE genes between C_3 species and C_4 species for following analysis.

We next investigated transcriptome-wide amino acid changes predicted from orthologues of C_3 and C_4 *Flaveria* species using the process shown in Fig. S5 (Supplementary Methods). To estimate the accuracy of the predicted peptide sequences from our results, we conducted a comparative study of protein sequences from UniProtKB (http://www.uniprot.org) with those from our results, we found that our predicted peptide sequences are as good as, if not better than, the sequences from UniProtKB in terms of accuracy (details see Supplementary Results and Table S3). As a result, we obtained 1,018 genes encoding at least one amino acid change between C_3 and C_4 *Flaveria* species. 56 out of these 1,018 genes also showed significantly differentially expression between C_3 and C_4 species, which was termed as the

142 modified genes.

143 **The modified genes showed coordinated and abrupt changes along the C_4 evolutionary pathway in**

144 **the genus *Flaveria***

145 Genes related to C_4 photosynthesis pathway and genes related to cyclic electron transport (CET)

146 chain were significantly enriched in the 56 modified genes. ("BH" correlated $P<0.05$, *Fisher's* exact test).

147 We systematically discussed these genes and their changes during C_4 evolution in *Flaveria* with gene

148 expression and predicted protein sequences.

149 *Genes encoding proteins associated with the C_4 pathway*

150 Nine genes encoding proteins associated with the C_4 pathway were identified, including those

151 encoding three C_4 cycle enzymes, PEPC, PPDK and NADP-ME, two regulatory proteins, PPDK

152 regulatory protein (PPDK-RP) and PEPC protein kinase A (PPCKA), two aminotransferases, Alanine

153 aminotransferase (AlaAT) and aspartate aminotransferase 5 (AspAT5), and two transporters, BASS2 and

154 sodium: hydrogen (Na^+/H^+) antiporter 1 (NHD1) (Table 1). In terms of protein sequence, the major

155 predicted amino acid changes in C_4 species occurred at N7 for all of the nine genes (Fig. 2, Figs. S6,

156 Table 1). For example, PEPC in the C_4 *Flaveria* species had 41 predicted amino acid changes compared

157 with those in the C_3 species, which were mapped onto the *Flaveria* phylogeny determined by Lyu *et al.*

158 [22]. One of the predicted changes occurred at N6 (D396 in C_4 species, hereafter D396), and 34 occurred

159 at N7 (Fig. 2A). The six other predicted amino acid changes occurred at N7 or after N7, although the

160 incomplete assembly of PEPC transcripts from *F. palmeri* and *F. vaginata* did not allow resolution of the

161 predicted amino acid sequences. These results suggest an evolutionary jump in the protein sequence at

162 N7 for C_4 enzymes.

In terms of gene expression, all the nine genes showed higher transcript abundance in C_4 species than in C_3 species and a comparable level in C_4-like and C_4 species (Table 1). To calculate the relative gene expression changes of each ancestral node, the FPKM values of each ancestral node were predicted and the relative difference were calculated (see Methods). In general, C_4 species showed a 7.6-fold to 123.6-fold of FPKM values compared with C_3 species. Similar to the pattern of changes for protein sequences, seven of the nine genes showed that the biggest relative changes of gene expression at N7. Whereas, both NADP-ME and AlaAT showed the biggest relative changes at two nodes of N3 and N6 with comparable levels (Fig. 2C and Fig. S6A). Our results hence suggested that the genes encoding proteins associated with C_4 pathway showed highly coordinated modification patterns in protein sequence and gene expression at N3, N6 and N7 during the evolutionary pathway of C_4 photosynthesis, while the majority of the predicted amino acid changes occurs at the N7.

Genes encoding proteins involved in CET chain

We identified genes encoding nine proteins that function in the CET chain, namely, proton gradient regulation 5 like (PGR5-like), the chloroplast NAD(P)H dehydrogenase complex (Ndh) L2-2 (NdhL2-2), NdhV, Ndh18, NdhU, NdhM, Ndh48, NdhB4, and chlororespiratory reduction 1 (CRR1). The transcripts encoding all the nine proteins showed higher abundances in C_4 species than C_3 species (Figs, 3 and Table 1). Compared to the genes encoded protein in C_4 pathway, the genes encoding proteins involved in CET chain showed the biggest changes at disperse nodes instead of at a single node. Specifically, the major changes of predicted protein sequences occurred at N6 and N7 and that of FPKM occurred at N3, N7 and N8. (Table1). Besides, the modification of protein sequence and FPKM is less coordinated in genes involved in CET chain, for example, the major change of PGR5-like occurred at N6 in predicted protein sequence and at N7 in FPKM (Fig. 3A).

185 *Genes encoding proteins in the photorespiratory pathway*

186 The establishment of photorespiratory pump (C_2 photosynthesis) is reported to be a prerequisite for
187 the evolution of C_4 photosynthesis based on theoretical modeling [25]. Therefore, we also investigate
188 genes involved in photorespiratory pathway in predicted protein sequence and FPKM.

189 One protein involved in photorespiration was included in the 56 modified genes, namely, glutamine
190 synthetase-like 1 (GSL1). Moreover, four other proteins in this pathway showed abundant amino acid
191 changes between C_3 and C_4 species, namely, glycine decarboxylase complex (GDC) H subunit (GDC-H),
192 serine hydroxymethyltransferase (SHM), glycerate kinase (GLYK), glutamine synthetase and glutamine
193 oxoglutarate aminotransferase (GOGAT) (Figs. 3, Table 1). In general, the predicted amino acid
194 substitution patterns of these five proteins were similar to those observed in the above described proteins
195 in C_4 pathways, with the major predicted amino acid changes in C_4 species occurring at N7 (Figs. 4,
196 Table 1), *e.g.,* 16 of 18 in GOGAT occurred at N7 (Fig. 4D). Generally, proteins in the photorespiratory
197 pathway showed fewer predicted amino acid changes than those in the C_4 pathway.

198 The abundance of transcripts encoding these five photorespiratory enzymes was comparable to
199 those in C_3 and C_3-C_4 species, and higher than that in the C_4 species (Figs. 4 A–E). When compared with
200 genes encoding C_4 pathway proteins, those encoding photorespiratory proteins showed larger differences
201 between C_4-like and C_4 species in clade A in terms of gene transcript abundance and protein sequence.
202 The greatest reduction of FPKM in these five genes was observed at N7 (Figs. 4, Table 1). Thus, this
203 suggested that the genes encoded proteins associated with photorespiratory pathway also showed
204 coordinated changes in protein sequence and gene expression during the evolutionary pathway of C_4
205 photosynthesis, and with the largest number of changes occurring at N7.

Physiological and anatomical characteristics related to C_4 photosynthesis show coordinated changes along the C_4 evolutionary pathway in *Flaveria*

To investigate whether C_4 related physiological characteristics also underwent coordinated changes during the evolution of C_4 photosynthesis in *Flaveria*, physiological characteristics taken from the literature [20, 35, 43] were mapped onto the *Flaveria* phylogeny (Fig. 5). The results revealed a step-wise change for most of the characteristics along the phylogenetic tree as previously suggested [20, 27, 35, 43] (Fig. 5); however, coordinated and abrupt changes were observed for a number of features. A major change in CO_2 compensation point (Γ) in *Flaveria* was first seen at N3, where the most ancestral C_3-C_4 species, *F. sonorensis*, was emerged which showed a decrease in Γ from 62.1 μbar of its closest C_3 relative *F. robusta* to 29.6 μbar (Fig. 5). The greatest changes in Γ in clade A occurred at N6, which showed a decrease in Γ from 24.1 μbar in *F. angustifolia* (C_3-C_4) to 9.0 μbar in *F. ramosissima* (C_3-C_4), followed by N7, where a decrease in Γ from 9.0 μbar in *F. ramosissima* (C_3-C_4) to 4.7 μbar in *F. palmeri* (C_4-like) was seen. The greatest decrease of Γ in clade B was observed between the two C_3-C_4 species, *F. floridana* and *F. chloraefolia* (C_3-C_4), where there was a decrease from 29 μbar to 9.5 μbar (Fig. 5). For photosynthetic water using efficiency (PWUE), photosynthetic nitrogen using efficiency (PNUE) and the slope of the response of the net CO_2 assimilation rate (A) versus Rubisco, the biggest changes occurred at N7 with increases of around 2-fold. In contrast, the percentage of ^{14}C fixed into four carbon acids showed no clear trend along the phylogenetic tree, although 3.91-fold and 1.76-fold increases were seen at N6 and N7, respectively. Interestingly, changes in all of these traits uniformly occurred at *F. brownii* in clade B, the only C_4-like species within this clade. Consequently, those data suggest that although there were gradual changes in physiological features along the C_3, C_3-C_4, C_4-like and C_4

227 trajectory, there are apparent jumps at N3, N6 and N7 in these physiological traits along the *Flaveria*
228 phylogeny (Fig. 5).

229 Anatomical traits [27, 31] were mapped onto the *Flaveria* phylogeny to investigate how these
230 features were modified along the evolution of C_4 (Fig. 5). For both the area of MCs and the ratio of the
231 area of MCs to that of BSCs (M: BS), the greatest modifications along the phylogeny were found
232 between *F. brownii* (C_4-like) and *F. floridana* (C_3-C_4), with a similar degree of change for both
233 characteristics (2.7-fold, Fig. 5). Anatomical data for *F. palmeri* (C_4-like) in clade A are not available,
234 however, large differences in anatomical features were found between the C_4-like *F. vaginata* and C_3-C_4
235 *F. ramosissima* [27]. The modification of M area first occurred at N2 which showed a 1.9-fold difference
236 between *F. robusta and F. cronquistii* followed by a 2.1-fold of difference between *F. ramosissima* and *F.*
237 *vaginata*. The major modification of the ratio of M and BS occurred at N2 with a 2.4-fold difference and
238 followed by N6 with a 1.6-fold difference and N7 with a 2-fold difference. Therefore, similar to the
239 evolutionary pattern of physiological features, large changes in anatomical features also emerged at N3,
240 N6 and the transition between C_3-C_4 species and C_4-like species. Interestingly, the ultrastructure of BSCs
241 chloroplasts showed an abrupt change at N7, with a dramatic decrease in grana thylakoid length and an
242 increase in stroma thylakoid length, whereas these features were comparable in the species at the base of
243 tree and in clade B [31].

244 **Coordinated change of protein sequence, gene expression and morphology with an evolutionary**
245 **jump at the transition between C_3-C_4 and C_4-like species along the species evolution**

246 Our above analysis showed that C_4 related genes and morphological features showed coordinated
247 changes with an obvious abrupt change at N7. Next, we asked whether species evolution also showed

evolutionary coordination and jump(s) along the species evolution in protein sequence, gene expression and morphology. To answer this question, we calculated the divergence matrics for protein sequence, gene expression, and morphological features between *F. cronquistii* (at the most basal place in the *Flaveria* phylogenetic tree) and other *Flaveria* species. The protein divergence was calculated as the rate of non-synonymous substitutions (dN) of all the genes that were used to construct the *Flaveria* phylogenetic tree from [22], the expression divergence was calculated as Euclidean distance of total expressed genes (see Methods), and the morphology divergence was calculated as Euclidean distance using previously coded morphology value from [28], which includes 30 types of morphology traits, such as life history, leaf shape, head types. Our result showed a high linear correlation between the protein divergence, gene expression divergence and morphology divergence, in particular between gene expression divergence and morphology divergence (R^2=0.9) (Figs. 6A-C). Thus, our results suggest a coordinated evolution of protein sequence, gene expression and morphology during the evolution. The linear correlation of gene expression divergence vs morphology divergence and protein divergence vs morphology divergence reflects that both gene expression changes and protein sequence changes are related to morphological changes during C_4 evolution. [2, 46]. Moreover, gene expression changes may be more directly related to morphology change than protein sequence changes [11]. It is likely that changes of developmental programs might be mainly due to changes in gene expression levels while changes in the protein sequences might contribute more to changes in metabolism.

Next, we predicted the protein sequence, transcript abundance and coded morphology value of ancestral nodes, which were then used to calculate the relative change of the three parameters at each node (see Methods). Surprisingly, protein sequence and gene expression showed significantly more changes at N7 than changes at other nodes (P<0.001, Tukey's test, "BH" adjusted， the same as

following), and the morphology showed the most changes at N7 with a marginal significant *P* value (*P*=0.06) (Fig. 6D), implying an evolutionary coordination and jump on whole transcript level also occurred in species evolution.

Discussion

Evolutionary coordination of different features implies a purifying selection towards a functional C_4 metabolism

Compared with C_3 photosynthesis, C_4 photosynthesis acquired many new features in gene expression, protein sequence, morphology and physiology (Figs. 2-5) [39]. We interpret these coordinated changes as a result of a strong purifying selection at this step. This is because though C_4 photosynthesis can gain higher photosynthetic energy conversion efficiency, highly specialized leaf and cellular anatomical features and biochemical properties of the involved enzymes are required. For example, increased cell wall thickness at the bundle sheath cell and decreased sensitivity of PEPC to malate inhibition are needed for C_4 plants to gain higher photosynthetic rates [44, 47]. Furthermore, to gain higher photosynthetic efficiency in C_4 plants, the ratio of the quantities of Rubisco content in BSCs and MCs is also critical [45]. In theory, if the C_4 decarboxlyation's inplace occurs before all other accompanying features, leaves will experience high leakage, *i.e.*, costing ATP for a futile cycle without benefit to CO_2 fixation. This will inevitably lead to lower quantum yield and a potential driving force for purifying selection. Further evidence for possible purifying selection comes from the observation that genes with cell-specific expression, such as PEPC, PPDK, and NADP-ME, displayed more changes in their predicted protein sequences than ubiquitously expressed genes, such as NDH components (Table 1,

290 Additional file 3). This is because, as discussed earlier, the redox environments between BSCs and MCs
291 might have changed dramatically during the completion of the C_4 cycle, with one of the most likely
292 change being having a more acidic environment due to increased production of Oxaloacetic acid (OAA)
293 and malate. Under such conditions, it is required for enzymes to alter their amino acid sequences to adapt
294 to the new cellular environments. The concurrent changes between gene expression divergence and
295 protein sequence divergence has also been demonstrated previously in animals [17, 46], which has been
296 similarly proposed to reflect negative selection for the involved genes [46].

297 The identified changes in the protein sequences, including amino acid changes, insertions, and
298 deletions (Table 1, Figs. 2-4), may enable the enzymes or proteins to improve biochemical and regulatory
299 properties to meet the demands of an altered cellular environment, for example, the increased fluxes
300 through the C_4 cycle [42]. It is worth to mention that some of these predicted amino acid changes have
301 been reported to be functional, such as the S774 and G884 residues in C_4 PEPC determines the high
302 substrate affinity and low inhibitor affinity of this enzyme, respectively. [7, 33]. Besides, many of the
303 predicted amino acid changes involve in residues that can be post-translationally modified, for example,
304 six residues in PPDK changed to Serine (S) in C_4 species, which can all be target for phosphorylation and
305 hence functional modification.

306 We also investigated whether these revised amino acid residues in C_4 species are under positive
307 selection. We used the "branch-site" model of PAML 4.8 to infer the positively selected amino acid sites
308 for genes in Table 1 (Supplemental Method). However, we found a limited number of sites showing
309 positive selection with a threshold of posterior probability (PP) based on Bayes Empirical Bayes being
310 higher than 0.95, *e.g.*, 147S of NADP-ME (PP: 0.953) and 147S of PPCKA (PP: 0.996). The 508V of

311 PEPC showed a significant PP of 0.972, suggesting a potential role of this site during C4 evolution.
312 Interestingly, two known functional sites of PEPC, namely, 774A and 884S, did not show PP higher than
313 0.95, (774A showed a PP of 0.867 and 884S showed a PP of 0.859). We emphasize here that in this study,
314 for a given gene, if complete protein sequence based on de novel assembly are not available for a species,
315 this species was excluded during prediction of positive selected sites. This exclusion inevitably decreases
316 the power of positive selection detection. From this perspective, all those identified sites under position
317 selection have a high chance of bearing biological function during C_4 evolution, which need
318 experimental studies.

319 **Evolutionary jumps along the C_4 evolution in the *Flaveria* genus**

320 Among the nodes leading to the C_4 emergence in clade A, the N7, which is the most recent common
321 ancestor of C_4-like and C_4 species in clade A, shows the biggest change in protein sequence, gene
322 expression and morphology in both C_4 specific features and also non-C_4 specific features (Table 1, Figs.
323 2-6). There are also apparent changes in these features at N3 and N6. These three nodes reflect three
324 critical stages along the emergence of C_4 metabolism. First, at N3, there was a large degree of changes in
325 gene expression, protein sequence and morpholog . One of the most important events during this phase is
326 the re-location of GDC from MSCs to BSCs based on earlier western blot data [30, 38]. Here we found
327 that SHM showed decreased expression while most of other photorespiratory related enzymes showed
328 little changes (Figs. 4). Similarly, at this step, the majority of the C_4 related genes showed little changes
329 (Figs. 2). However, modification on gene expression and protein sequence on transcriptome level and
330 no-C_4 morphology features suggest that there is large number of changes at N3 (Figs. 6), and there is also
331 great decrease of CO_2 compensation point at this stage (Fig. 5).

N6 is the stage where we found the third largest degree of changes occurred in C_4 related features. At this stage, we observed large increase in transcript abundance in C_4 genes (Figs. 2 & Fig. S5) and photorespiratory genes (Figs. 4), and a dramatic increase in the percentage of ^{14}C incorporated into the four carbon acids occurred (Fig. 5). The modification of photorespiratory genes might be related to the optimization of C_2 cycle to decrease CO_2 concentrating point, which can increase fitness of plants under conditions favoring photorespiration [40]. The concurrent modification of C_4 enzyme, *i.e.,* PEPC, NADP-ME, PEPCKA, *etc*, which are also involved in nitrogen rebalancing, is consistent with the notation that C_4 cycle might be evolved as a result of rebalancing nitrogen metabolism after GDC moving from MC to BSC [24]. The fact that there is little change in the $\delta^{13}C$ in the C_3-C_4 intermediate as compared to that of C_3 species suggests that the contribution of CO_2 fixation following C_4 pathway is relatively minor, *i.e.*, less than 15%, estimated based on an $\delta^{13}C$ value of -27.6 in *F. ramosissima* (Fig. 5), again supporting the initial role of increased C_4 enzymes is not for enhancing CO_2 fixation. It is worth pointing out here that the measured initial carbon fixation in the form of C_4 compound was 46% (Fig. 5), higher than those estimated based on the $\delta^{13}C$ value. This is possibly because though malate releases CO_2 into BSCs as a result of the nitrogen rebalancing pathway, most of the CO_2 was not fixed by Rubisco, either due to lack of sufficient Rubisco activity in BSCs or due to lack of required low BSCs cell wall permeability to maintaining high CO_2 concentration in BSCs *etc*.

N7 witnesses abrupt changes for both the gene expression and proteins sequence and morphology (Figs. 2-6). The majority of the C_4 related genes showed the most modification in gene expression and protein sequence at N7, especially for genes in C_4 cycle and photorespiratory pathway. Moreover, N7, where C_4-like species (clade A) appear, represents a dramatic shift of CO_2 fixation from being dominated by a C_2 concentrating mechanism to being dominated by a C_4 concentrating mechanism. Based on the

354 $\delta^{13}C$ value in *F. palmeri*, the fixation through the C_4 concentrating mechanism is up to 93%, which is
355 consistent with the measured proportion of initial carbon fixation in the form of C_4, compound (Fig. 5),
356 suggesting at this step, the released CO_2 in the BSCs can be largely fixed by Rubisco. Whereas, the
357 transition between C_4-like to C_4 process is an evolutionarily "down-hill" process and most of
358 optimization occurred through fine-tuning expression abundance.

359 ## Materials and Methods

360 ### Data retrieval

361 RNA-Seq data of *Flaveria* species were downloaded from the Sequence Read Achieve (SRA) of
362 the National Center for Biotechnology Information (NCBI) (Supplementary Methods). All accession
363 numbers for RNA-Seq data are shown in Table S1.

364 CO_2 compensation points (Γ) (except for *F. kochiana*), $\delta^{13}C$ (except for *F.kochiana*), $\%O_2$ inhibition
365 of P_{max} (except *F. kochiana*), and CO_2 assimilation rates were from [20]. Γ, δ^{13} and $\%O_2$ inhibition of *F.*
366 *kochiana* were from [43]. Data for % initial C_4 products in total fixed carbon were from [29] . Data for
367 PWUE, PNUE, and net CO_2 assimilation rate (*A*) versus Rubisco content were from [43]. Data for M
368 area, M:BS ratio, vein density and number of ground tissue layers were from [27]. The values of M area,
369 M:BS ratio and vein density were measured from figures in McKown and Dengler [27] with GetData
370 (http://www.getdata-graph-digitizer.com). Mean values from 20 measurements were used.
371 Ultrastructural data of BS cell chloroplasts were from Nakamura *et al*. [31].

372 ### Transcriptome assembly and quantification

373 Transcripts of *Flaveria* species generated with Illumina sequencing were assembled using Trinity

374 (version 2.02) [10] with default parameters (Table S1). Contigs of four *Flaveria* species from 454
375 sequencing data were assembled using CAP3 [16] with default parameters. In all cases, only contigs of at
376 least 300 bp in length were saved. Transcript abundances of 31 *Flaveria* samples were analyzed by
377 mapping Illumina short reads to assembled contigs of corresponding species and then normalized to the
378 fragment per kilobase of transcript per million mapped reads (FPKM) using the RSEM package (version
379 1.2.10) [21]. Functional annotations of *Flaveria* transcripts were determined by searching for the best hit
380 in the coding sequence (CDS) dataset of *Arabidopsis thaliana* (Arabidopsis) in TAIR 10
381 (http://www.arabidopsis.org) by using BLAST in protein space with E-value threshold 0.001. If multiple
382 contigs shared the same best hit in CDS reference of Arabidopsis, then the sum FPKM of those contigs
383 was assigned to the FPKM value of the gene in *Flaveria*.

384 To estimate the consistence of *Flaveria* gene annotation, we used OrthoFinder[6] to predict the
385 orthologous group based on the annotated *Flaveria* gene together with gene of Arabidopsis from TAIR10,
386 and then calculated the consistence between gene annotation and orthologous group in two ways. (1) If
387 the orthologous group contains Arabidopsis gene(s), the consistency was calculated as the percentage of
388 genes that have the same annotation with the Arabidopsis gene(s). (2) If there the orthologous group does
389 not contain Arabidopsis gene, we calculated the percentage of genes for each gene ID in this group, and
390 the highest percentage was assigned to the consistency.

391 To make the FPKM comparable across different samples, we normalized the FPKM value by a
392 scaling strategy as used by Brawand *et al*. [2]. Specifically, among the transcripts with FPKM values
393 ranking in 20%-80% region in each sample, we identified the 1000 genes that had the most-conserved
394 ranks among 29 leaf samples, which were then used as an internal reference, and the transcript of each

sample was normalized according to the mean value of these 1000 genes in the sample. We then multiplied all the FPKM values in all samples by the mean value of 1000 genes in the 29 leaf samples. The three samples from C_3 species and eight samples from C_4 species (Table S1) were used to recall differentially expressed (DE) genes applying edgeR [34], and the Benjamini-Hochberg ("BH") procedure was used in multiple testing correction with a threshold of *P* ("BH" corrected) to be 0.05.

Investigation the species used in this study being from hybrid of two species

To investigate whether the intermediate species used in this study are from hybrid offspring of two species, DNA sites that expressed different alleles were identified, which termed as mixed sited. The mixed sites were identified based on RNA-Seq data as described in [22]. Hybrid offspring from two different species are expected to have higher percentage of mixed sites among all expressed sites than no hybrid species. To create a positive background of hybrid samples, RNA-Seq data of 16 species were pair-wisely mixed and their mixed sites were also identified. The mixed sites of the known hybrid sample *F. pringlei** originated from *F. pringlei* ×*F. angustiflolia* in [22] were also identified. The percentage of mixed sites was calculated as the ratio of mixed sites to the total expressed DNA sites in a certain sample.

Protein divergence, gene expression divergence and morphology divergence

Pair-wise protein divergence (dN) was calculated by applying codeml program in PAML package [49] by using F3X4 condon frequency. The input super CDS sequence was from the linked coding sequences (CDS) as used in construct phylogenetic tree of *Flaveria* genus [22], which contains 2462 genes. Gene expression divergence was calculated as Euclidean distance applying R package based on gene expression values (FPKM) of 1,2218 genes. Encoded morphology values of 30 morphology traits

were from [28]. The morphology divergence was calculated as Euclidean distance of morphology values. Expression and morphology values were normalized using quantile normalization applying preprocessCore package in R. Linear regression of pair-wise correlation was inferred apply lm function in R package.

Relative difference of each ancestral node in the phylogenetic tree

The morphological characteristics, gene expression abundance, and protein sequences at the whole transcriptomic scale were predicted using FASTML [1]. The protein alignment was from [22]. Gene expression abundance and morphological characteristics of all ancestral nodes were predicted by applying ape package of R which uses a maximal likelihood method. For all C_4 related gene expression, protein sequences and physiological data, their values of the ancestral nodes were assigned to those of the most recent species derived from the node.

Relative difference of protein sequence at each ancestral node was inferred by comparing the sequence at this node (N) with the nearest preceding node of N (N[pre]), *e.g.*, the number of different amino acid between N2 with N1 is the number of changed amino acid at N2. The number of different amino acid changes divided by the aligned length of the protein was calculated as relative protein difference for each gene. Relative difference of gene expression and morphology were calculated as (N -N[pre])/N[pre]. In most cases, the nearest preceding node of N[i] is N[i-1], there are two exceptions: the ancestral node of N11 is N5, and N10 is N8. One-way ANOVA analysis followed by Tukey's Post Hoc test was used to calculate the significance of relative difference between any two ancestral nodes. *P* values were adjusted by *Benjamin-Hochberg* (BH) correction.

List of Abbrevations

A: CO2 assimilation rate; AlaAT: Alanine aminotransferase; AspAT5: aspartate aminotransferase 5; BSCs: bundle sheath cells; CET: cyclic electron transport; CRR1: chlororespiratory reduction 1; DE: differentially expressed; FPKM: fragments per kilobase of transcript per million mapped reads; GDC: glycine decarboxylase complex; GLYK: glycerate kinase; GOGAT: glutamine synthetase and glutamine oxoglutarate aminotransferase; GSL1: glutamine synthetase-like 1; MCs: mesophyll cells; NADP-ME: NADP-dependent malic enzyme; NCBI: National Center for Biotechnology Information; Ndh: NADH dehydrogenase-like; NHD1: sodium: hydrogen (Na+/H+) antiporter 1; PEPC: phosphoenolpyruvate carboxylase; PGR5-like: proton gradient regulation 5 like; PIFI : post-illumina chlorophyll fluorescence increase; PNUE: instantaneous photosynthetic nitrogen use efficiency; PPCKA: PEPC protein kinase A; PPDK-RP: PPDK regulatory protein; PPDK: pyruvate, orthophosphate dikinase; PWUE: instantaneous photosynthetic water use efficiency; SHM: hydroxymethyltransferase; SRA: Sequence Read Achieve; Γ: CO_2 compensation point;

Availability of data and materials

All data generated during this study are included in this published article and its supplementary information files. Codes used during the current study are available from the corresponding author on reasonable request.

Acknowledgements

The authors thank Haiyang Hu and Yimin Tao for great discussion and suggestion; we also thank LetPub (www.letpub.com) for its linguistic assistance during the preparation of this manuscript. This work was sponsored by Shanghai Sailing Program [17YF421900], National Science Foundation of

China [31701139 to Ming-Ju Amy Lyu, 30970213 to Xin-Guang Zhu], and Bill & Melinda Gates Foundation [OPP1014417].

Authors' Contributions

MAL, UG, SC and HC conducted the analysis and wrote the paper. SK, JMH, RFS, ML, GKS, PW and XGZ designed the study and wrote the paper.

Compliance with ethical standards

Competing interests

The authors declare that they have no conflict of interest.

Reference

1. Ashkenazy H, Penn O, Doron-Faigenboim A, Cohen O, Cannarozzi G, Zomer O, Pupko T: FastML: a web server for probabilistic reconstruction of ancestral sequences. Nucleic Acids Res 40: W580-4 (2012).
2. Brawand D, Soumillon M, Necsulea A, Julien P, Csardi G, Harrigan P, Weier M, Liechti A, Aximu-Petri A, Kircher M, Albert FW, Zeller U, Khaitovich P, Grutzner F, Bergmann S, Nielsen R, Paabo S, Kaessmann H: The evolution of gene expression levels in mammalian organs. Nature 478: 343-8 (2011).
3. Brown NJ, Parsley K, Hibberd JM: The future of C4 research--maize, Flaveria or Cleome? Trends Plant Sci 10: 215-21 (2005).
4. Dengler N, Nelson T: Leaf structure and development in C4 plants. Sage, R, F,, Monson, R, K ed (s). C4 plant biology.. Academic Press: San Diego, etc (1999).
5. Edwards GE, Ku MS: Biochemistry of C3-C4 intermediates. In: Hatch MD, Boardman NK, editors. The biochemistry of plants. New York: Academic Press: 275-325 (1987).
6. Emms DM, Kelly S: OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol 16: 157 (2015).
7. Engelmann S, Blasing OE, Gowik U, Svensson P, Westhoff P: Molecular evolution of C4 phosphoenolpyruvate carboxylase in the genus Flaveria--a gradual increase from C3 to C4 characteristics. Planta 217: 717-25 (2003).
8. Engelmann S, Wiludda C, Burscheidt J, Gowik U, Schlue U, Koczor M, Streubel M, Cossu R, Bauwe H, Westhoff P: The gene for the P-subunit of glycine decarboxylase from the C4 species Flaveria trinervia: analysis of transcriptional control in transgenic Flaveria bidentis (C4) and Arabidopsis (C3). Plant Physiol 146: 1773-85 (2008).
9. Gowik U, Brautigam A, Weber KL, Weber AP, Westhoff P: Evolution of C4 photosynthesis in the genus Flaveria: how many and which genes does it take to make C4? Plant Cell 23: 2087-105 (2011).
10. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A: Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29: 644-52 (2011).
11. Hart JC, Ellis NA, Eisen MB, Miller CT: Convergent evolution of gene expression in two

high-toothed stickleback populations. PLoS Genet 14: e1007443 (2018).

12. Hatch MD: C4 photosynthesis: a unique blend of modified biochemistry, anatomy and ultrastructure. Biochimica et Biophysica Acta 895: 81-106 (1987).

13. Hatch MD, Osmond CB: Compartmentation and transport in C4 photosynthesis. Encyclopedia of Plant Physiology 3: 144-184 (1976).

14. Hatch MD, Slack CR: A new enzyme for the interconversion of pyruvate and phosphopyruvate and its role in the C4 dicarboxylic acid pathway of photosynthesis. Biochem J 106: 141-6 (1968).

15. Huang X, Han B: Natural variations and genome-wide association studies in crop plants. Annu Rev Plant Biol 65: 531-51 (2014).

16. Huang X, Madan A: CAP3: A DNA sequence assembly program. Genome Res 9: 868-77 (1999).

17. Hunt BG, Ometto L, Keller L, Goodisman MAD: Evolution at Two Levels in Fire Ants: The Relationship between Patterns of Gene Expression and Protein Sequence Evolution. Molecular Biology and Evolution 30: 263-271 (2013).

18. Johnson HS, Hatch MD: The C4-dicarboxylic acid pathway of photosynthesis. Identification of intermediates and products and quantitative evidence for the route of carbon flow. Biochem J 114: 127-34 (1969).

19. Kadereit G, Bohley K, Lauterbach M, Tefarikis DT, Kadereit JW: C3 -C4 intermediates may be of hybrid origin - a reminder. New Phytol 215: 70-76 (2017).

20. Ku MS, Wu J, Dai Z, Scott RA, Chu C, Edwards GE: Photosynthetic and photorespiratory characteristics of flaveria species. Plant Physiol 96: 518-28 (1991).

21. Li B, Dewey CN: RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12: 323 (2011).

22. Lyu MJ, Gowik U, Kelly S, Covshoff S, Mallmann J, Westhoff P, Hibberd JM, Stata M, Sage RF, Lu H, Wei X, Wong GK, Zhu XG: RNA-Seq based phylogeny recapitulates previous phylogeny of the genus Flaveria (Asteraceae) with some modifications. BMC Evol Biol 15: 116 (2015).

23. Lyu MJ, Gowik U, Kelly S, Covshoff S, Mallmann J, Westhoff P, Hibberd JM, Stata M, Sage RF, Lu H, Wei X, Wong GK, Zhu XG: RNA-Seq based phylogeny recapitulates previous phylogeny of the genus Flaveria (Asteraceae) with some modifications. BMC Evolutionary Biology 15: 116 (2015).

24. Mallman J, Heckmann D, Brautigam A, Lercher MJ, Webb APM, Westhoff P, Gowik U: The

role of photorespiration during the evolution of C4 photosynthesis in the genus Flaveria. eLife 3: e02478 (2014).

25. Mallmann J, Heckmann D, Brautigam A, Lercher MJ, Weber AP, Westhoff P, Gowik U: The role of photorespiration during the evolution of C4 photosynthesis in the genus Flaveria. Elife 3: e02478 (2014).

26. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM: Finding the missing heritability of complex diseases. Nature 461: 747-53 (2009).

27. McKown AD, Dengler NG: Key innovations in the evolution of Kranz anatomy and C4 vein pattern in Flaveria (Asteraceae). Am J Bot 94: 382-99 (2007).

28. McKown AD, Moncalvo J-M, Dengler NG: Phylogeny of Flaveria (Asteraceae) and inference of C4 photosynthesis evolution. American Journal of Botany 92: 1911-1928 (2005).

29. Moore Bd, Ku M, S. B, Edwards G, E.: C4 photosynthesis and light-dependent accumulation of inorganic carbon in leaves of C3-C4 and C4 Flaveria species. Australian Journal of Plant Physiology 14: 658-668 (1987).

30. Morgan CL, Turner SR, Rawsthorne S: Coordination of the Cell-Specific Distribution of the 4 Subunits of Glycine Decarboxylase and of Serine Hydroxymethyltransferase in Leaves of C3-C4 Intermediate Species from Different Genera. Planta 190: 468-473 (1993).

31. Nakamura N, Iwano M, Havaux M, Yokota A, Munekage YN: Promotion of cyclic electron transport around photosystem I during the evolution of NADP-malic enzyme-type C4 photosynthesis in the genus Flaveria. New Phytol 199: 832-42 (2013).

32. Pattin KA, Moore JH: Genome-wide association studies for the identification of biomarkers in metabolic diseases. Expert Opin Med Diagn 4: 39-51 (2010).

33. Paulus JK, Schlieper D, Groth G: Greater efficiency of photosynthetic carbon fixation due to single amino-acid substitution. Nature Communications 4: 1518 (2013).

34. Robinson MD, McCarthy DJ, Smyth GK: edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26: 139-40 (2010).

35. Rumpho ME, Ku MS, Cheng SH, Edwards GE: Photosynthetic Characteristics of C(3)-C(4) Intermediate Flaveria Species : III. Reduction of Photorespiration by a Limited C(4) Pathway of Photosynthesis in Flaveria ramosissima. Plant Physiol 75: 993-6 (1984).

566 36. Sage RF: The evolution of C4 photosynthesis. New Phytologist 161: 341-347 (2003).
567 37. Sage RF, Christin PA, Edwards EJ: The C4 plant lineages of planet Earth. Journal of
568 Experimental Botany 62: 3155-3169 (2011).
569 38. Sage RF, Sage TL, Kocacinar F: Photorespiration and the evolution of C4 photosynthesis.
570 Annu Rev Plant Biol 63: 19-47 (2012).
571 39. Sage RF, Zhu X-G: Exploiting the engine of C4 photosynthesis. Journal of Experimental
572 Botany 62: 2989–3000 (2011).
573 40. Sage TL, Busch FA, Johnson DC, Friesen PC, Stinson CR, Stata M, Sultmanis S, Rahman BA,
574 Rawsthorne S, Sage RF: Initial events during the evolution of C4 photosynthesis in C3
575 species of Flaveria. Plant Physiol 163: 1266-76 (2013).
576 41. Slack CR, Hatch MD, Goodchild DJ: Distribution of enzymes in mesophyll and
577 parenchyma-sheath chloroplasts of maize leaves in relation to the C4-dicarboxylic acid
578 pathway of photosynthesis. Biochem J 114: 489-98 (1969).
579 42. Stitt M, Zhu X-G: The large pools of metabolites involved in intercellular metabolite
580 shuttles in C4 photosynthesis provide enormous flexibility and robustness in a
581 fluctuating light environment. Plant, Cell & Environment: n/a-n/a (2014).
582 43. Vogan PJ, Sage RF: Water-use efficiency and nitrogen-use efficiency of C(3) -C(4)
583 intermediate species of Flaveria Juss. (Asteraceae). Plant Cell Environ 34: 1415-30
584 (2011).
585 44. Wang Y, Long SP, Zhu XG: Elements required for an efficient NADP-malic enzyme type C4
586 photosynthesis. Plant Physiol 164: 2231-46 (2014).
587 45. Wang Y, Virtanen J, Xue Z, Zhang Y: I-TASSER-MR: automated molecular replacement for
588 distant-homology proteins using iterative fragment assembly and progressive sequence
589 truncation. Nucleic Acids Res 45: W429-W434 (2017).
590 46. Warnefors M, Kaessmann H: Evolution of the Correlation between Expression Divergence
591 and Protein Divergence in Mammals. Genome Biology and Evolution 5: 1324-1335
592 (2013).
593 47. Wedding RT, Black MK, Meyer CR: Inhibition of phosphoenolpyruvate carboxylase by
594 malate. Plant Physiol 92: 456-61 (1990).
595 48. Westhoff P, Gowik U: Evolution of c4 phosphoenolpyruvate carboxylase. Genes and
596 proteins: a case study with the genus Flaveria. Ann Bot 93: 13-23 (2004).
597 49. Yang Z: PAML: a program package for phylogenetic analysis by maximum likelihood.
598 Comput Appl Biosci 13: 555-6 (1997).

50. Zhu X-G, Shan L, Wang Y, Quick WP: C4 Rice - an ideal arena for systems biology research. Journal of Integrative Plant Biology 52: 762-770 (2010).

Supplementary Information:

Additional file 1: Includes supplemental methods, figures and tables

Additional file 2: The alignments of proteins

Additional file 3: FPKM of all genes; DE genes between C_3 vs C_4 species; DE genes between C_3 and type I C_3-C_4 from clade A; DE genes between C_3 vs all C_3-C_4 species and 56 genes that showed differential expression and at least one amino acid change between C_3 and C_4 species.

Figures

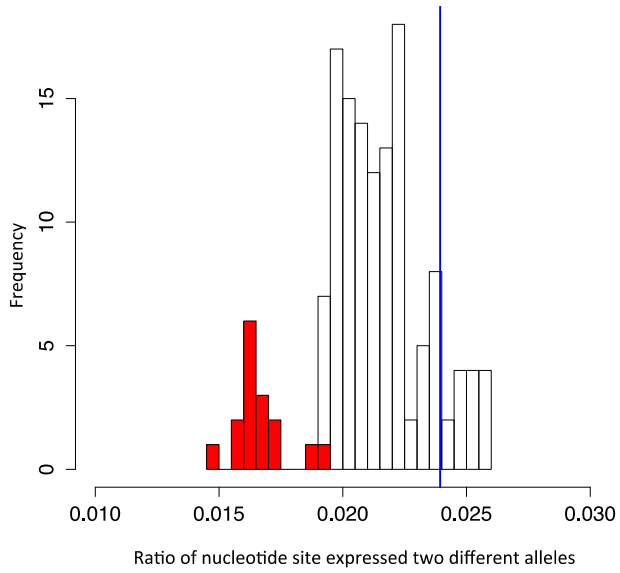

Figure 1. Estimation of the probability of RNA-Seq data from hybrid species

The bars show the distribution of ratio of nucleotide sites expressing two different alleles (mixed site). Mixed RNA-Seq samples are generated by pair-wise mixing RNA-Seq data of 16 *Flaveria* species, which mimics the case of hybridization. The ratio of mixed site in the mixed RNA-Seq samples is showed in grey bars (positive control). The ratio of mixed site of 16 *Flaveria* species is showed in red bars (real causes). The ratio of hybrid sample *F. pringlei**, was represented in blue vertical line.

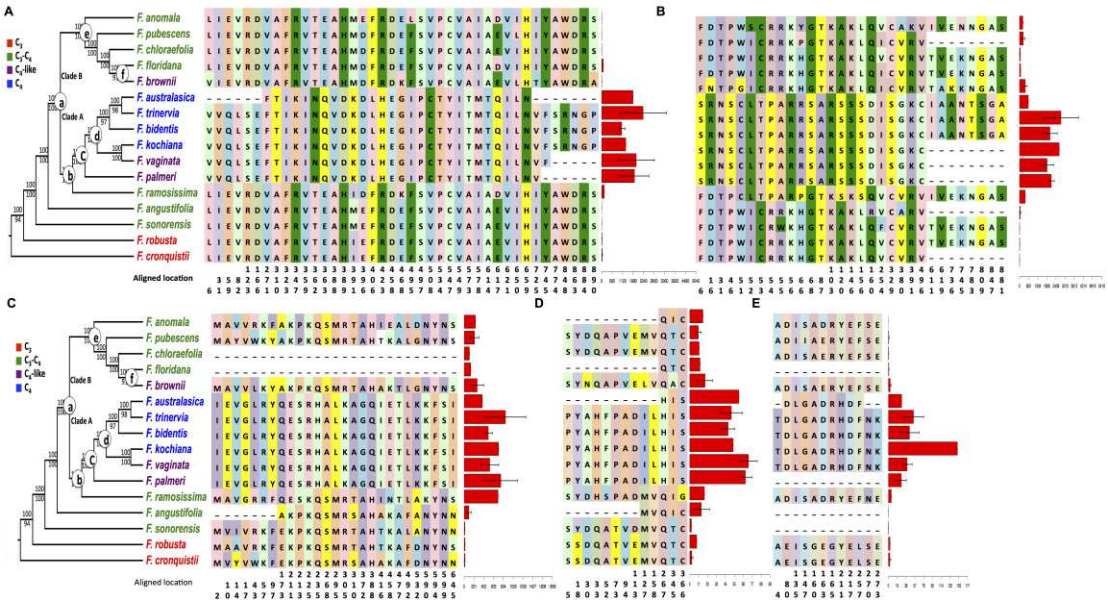

Figure 2. Modifications in genes in C_4 pathway in predicted protein sequences and transcript abundances mapped to the *Flaveria* phylogeny

The predicted amino acid changes and the transcript abundance (FPKM) of the genes encoding the enzymes in C_4 pathway are shown. Only the amino acid residues predicted to be different between C_3 and C_4 species are superimposed on the schema of *Flaveria* phylogenetic tree modified from Lyu *et al.*, 2015. The colors of amino acid residues have no meaning and are only for visualization purposes. Numbers below the amino acids indicate the location sites in the multiple sequence alignments. FPKM values with standard errors are shown to the right of the amino acid changes as red bars. A: phospho*enol*pyruvate carboxylase (PEPC); B: pyruvate orthophosphate dikinase (PPDK); C: NADP-malic enzyme (NADP-ME); D: pyruvate orthophosphate dikinase regulatory protein (PPDK-RP); E: phospho*enol*pyruvate protein kinase A (PPCKA). Protein sequences from UniprotKP are: *F. trinervia* PEPC, P30694; *F. bidentis* PPDK, Q39735; *F. brownii* PPDK, Q39734; and *F. trinervia* PPDK, P22221. Sequence alignments are available in Additional file 2.

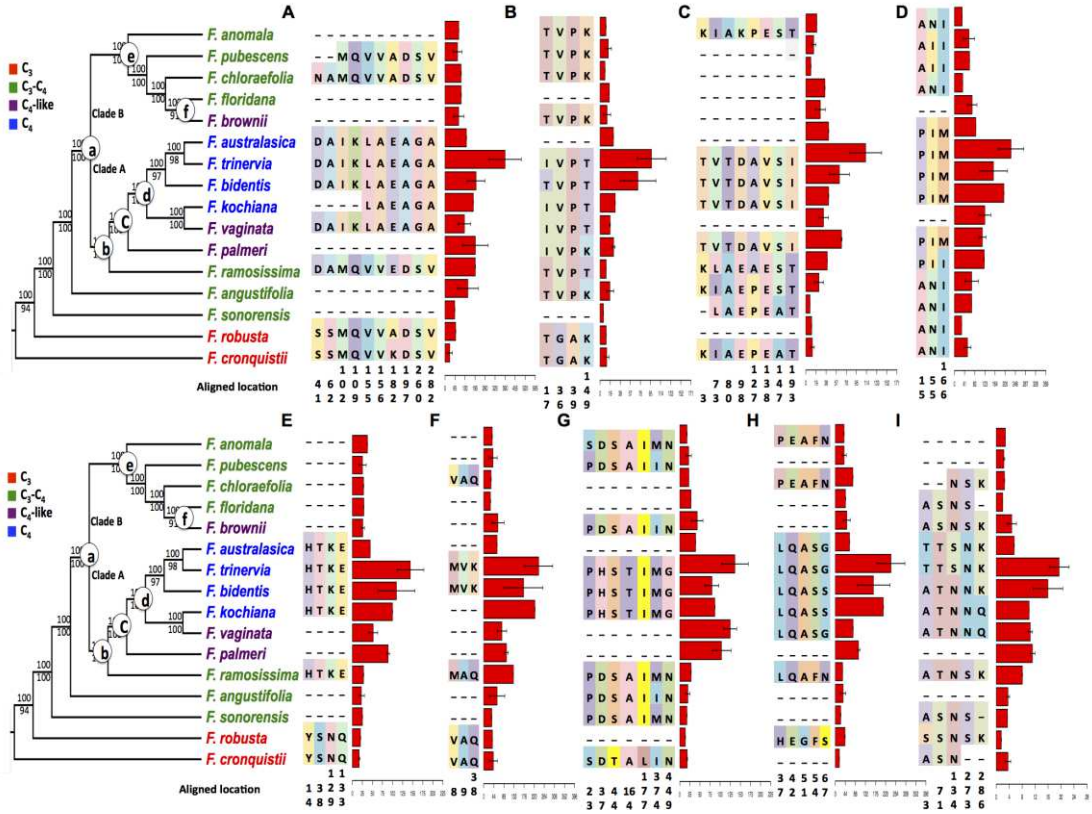

Figure 3. Modifications in the predicted amino acid sequences of proteins involved in cyclic electron transport and transcript abundances of the cognate transcripts mapped to the *Flaveria* phylogeny

Changes in predicted amino acid sequence in proteins involved in cyclic electron transport chain and abundances (FPKM) of their cognate transcripts in C_4 and C_3 *Flaveria* species are shown. Only the amino acid residues predicted to be different between C_3 and C_4 species are superimposed on the schema of *Flaveria* phylogenetic tree modified from Lyu *et al.*, 2015. The marked colors of amino acid residues have no meaning and are only for visualization purposes. Numbers below the amino acids indicate the location site in the multiple sequence alignments. FPKM values with standard errors are represented to the right of the amino acid changes as red bars. A: protein gradient regulation 5 like protein (PGR5-like); B: NADH dehydrogenase-like (Ndh) L2 subunit (Ndh L2); C: NdhV; D: Ndh16; E: NdhU; F: NdhM; G: Ndh48; H: NdhB4; I: chlororespiration reduction 1 (CRR1). The sequence alignments are available in Additional file 2.

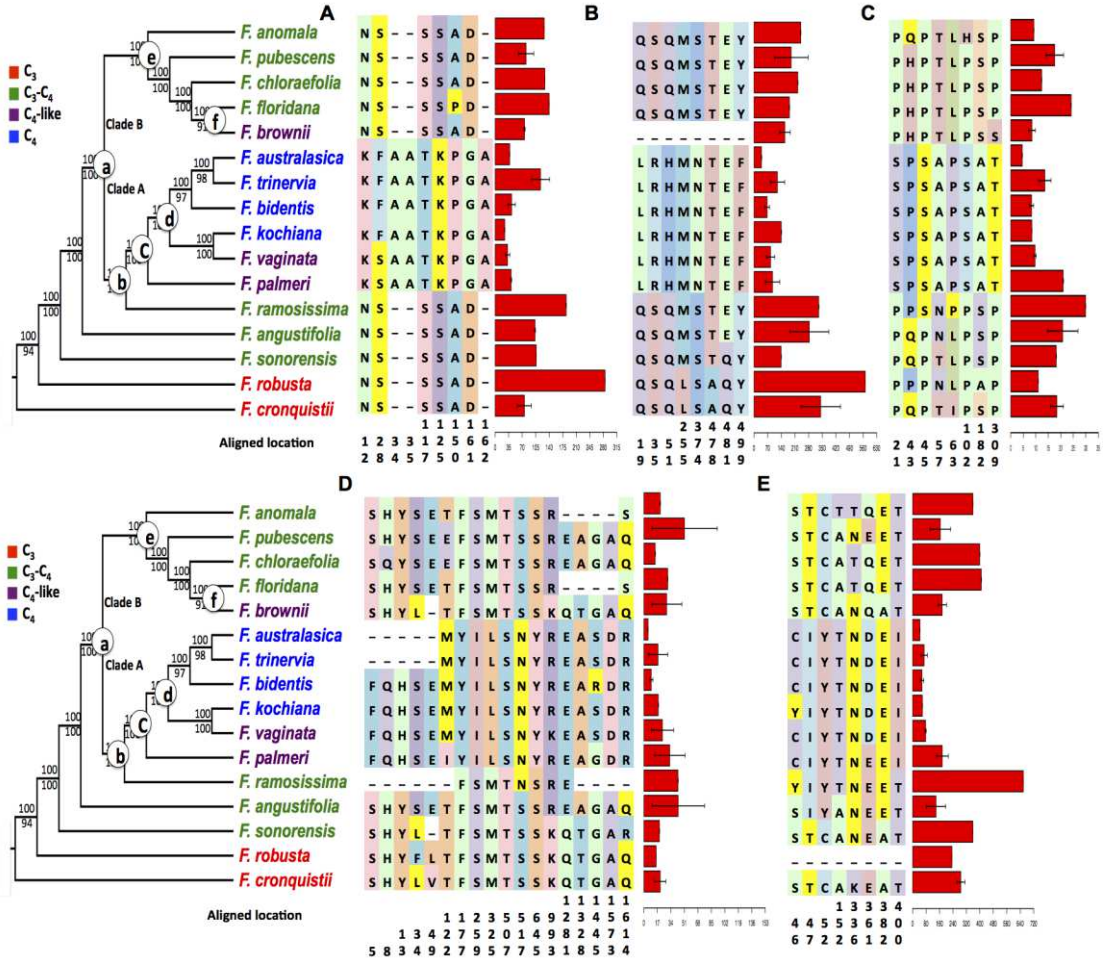

Figure 4. Modifications in photorespiratory protein predicted amino acid sequences and cognate transcript abundances mapped to the *Flaveria* phylogeny

The predicted amino acid changes in photorespiratory proteins between C_4 and C_3 *Flaveria* species and the transcript abundance (FPKM) of genes encoding the proteins are shown. Only the amino acid residues that are predicted to be different between C_3 and C_4 species are superimposed on the schema of *Flaveria* phylogenetic tree modified from Lyu *et al.*, 2015. The marked colors of amino acid residues have no meaning and are only for visualization purposes. Numbers below the amino acids indicate the location site in the multiple sequence alignments. FPKM values with standard errors are represented to the right of the amino acid changes as red bars. A, Glutamine synthetase like 1 (GSL1); B: glycine decarboxylase complex H subunit (GDC-H); C: serine hydroxymethyltransferase (SHM); D: glycerate kinase (GLYK); E: glutamine synthetase and glutamine oxoglutarate aminotransferase (GOGAT).

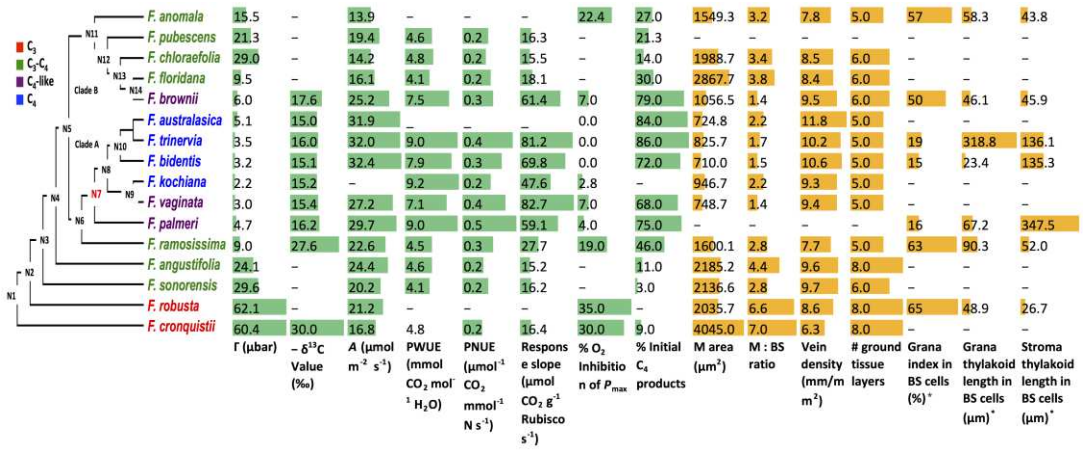

Figure 5. Changes in physiological and anatomical traits mapped onto the *Flaveria* phylogeny

Overall, C_4-related physiological (green and blue bars) and anatomical traits (orange and red bars) showed a step-wise change along the *Flaveria* phylogenetic tree; however, a number of the traits showed greater more significant changes at certain nodes. *Grana index: total length of grana/total length of thylakoid membrane X 100. (Abbreviations: Γ: CO_2 compensation point; *A*: CO_2 assimilation rate; PWUE: instantaneous photosynthetic water use efficiency; PNUE: instantaneous photosynthetic nitrogen use efficiency; response slope: slope of the response of net CO_2 assimilation rate versus leaf Rubisco content; M: mesophyll; BS: bundle sheath.) Data are from references as given in the Methods.

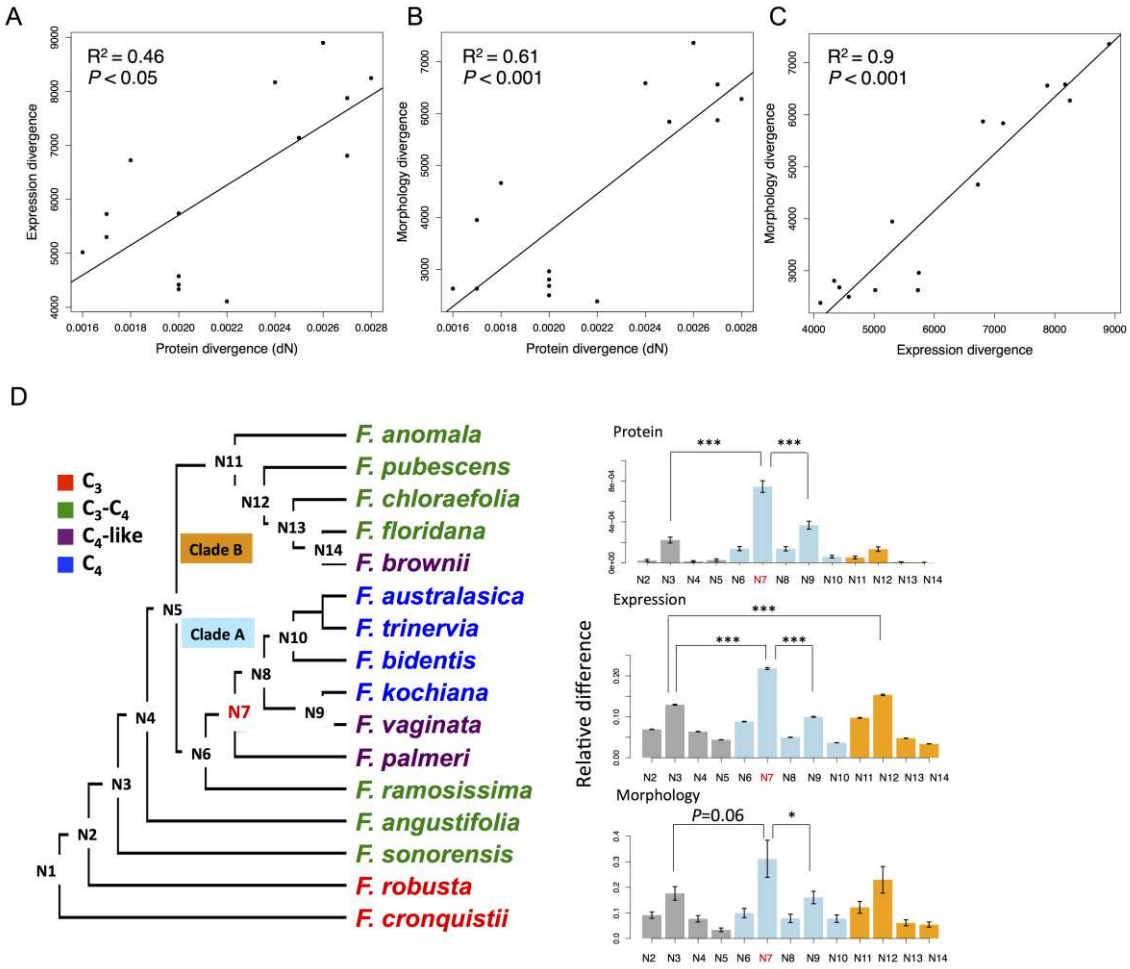

Figure 6. Coordinated evolution of protein sequence, gene expression and morphology with an obvious jump change

Significant linear correlation between protein divergence, gene expression divergence and morphology divergence were showed in (A-C). Protein divergence was calculated as non-synonymous mutation (dN). Expression divergence and morphology divergence were calculated as Euclidean distance based on quantile normalized FPKM values and coded morphology values from Mckown *et al.*, 2005, respectively. All the relative divergences were the divergence between *F. cronquistii* and other *Flaveria* species. (D) Shows the relative difference of each ancestral node compared with its earlier ancestral node in protein sequence, gene expression and morphology. The left panel shows the schema of *Flaveria* phylogenetic tree modified from Lyu, *et al*., 2015. Each ancestral node was numbered in the evolutionary sequential order. *P* values are from One-way ANOVA analysis followed by Tukey's Post Hoc test and adjusted by *Benjamin-Hochberg* correction. The significant levels are: *: $P<0.05$; **: $P<0.01$; ***: $P<0.001$. The bar colors in grey/blue/orange represent species from basal/clade A/clade B of phylogenetic tree, respectively.

Table

Table 1. Proteins showing differences in amino acid sequence between C_3 and C_4 *Flaveria* species and the relative changes in their cognate transcripts

Ortholog in *A. thaliana*	Genes encoding proteins involved in	Mean FPKM (C_4)/mean FPKM (C_3)	FDR(EdgeR)	Length in Fcro (Frob)[α]	Protein length in *A. thaliana* (aa)	aa changes						Stage of key change(s) in sequence	Stage of key change(s) in FPKM[β]
						total aa change(s)	before N 5	at N5	at N6	at N7	after N7		
Gene in C4 pathway													
AT3G14940	**PEPC**	85.58	2.78E-06	966	968	41			1	>=34		N7	N7
AT4G15530	**PPDK**	123.6	9.10E-09	958	963	31			2 + 6-aa REP	>=15		N7	N7
AT1G79750	**NADP-ME**	26.64	6.64E-08	647	646	27		1	8	18		N7	N3 and N6
AT4G21210	**PPDK-RP**	7.57	1.63E-03	402	403	13	1		4	7		N7	N7
AT3G04530	**PEPC-k**	88.78	2.93E-03	281	278	12	3		2	7		N7	N7
AT1G72330	**AlaAT**	9.63	1.57E-04	544	553	9			2	7		N7	N3 an N6
AT4G31990	**AspAT5**	36.67	4.34E-06	459	453	3			1	1	1	N7	N3 and N7
AT2G26900	**BASS2**	39.12	5.30E-07	415	409	14			2	12		N7	N7
AT3G19490	**NHD1**	51.19	8.49E-07	576	576	15			2	13		N7	N7
Gene related to electron transport chain													
AT4G22890	**PGR5-like**	7.1	4.70E-02	328	324	10+17-aa INS	1		2+17-aa INS	7		N6	N7
AT1G14150	**NdhL2/PnsL2**	3.71	2.13E-02	190	190	4	2		1	1		before N5	N7
AT2G04039	**NdhV**	9.17	1.73E-02	227	199	8		1	6	1		N6	N3
AT5G43750	**Ndh18/PnsB5**	6.8	6.27E-02	224	212	3			2	1		N6	N8
AT5G21430	**NdhU/CRRL**	8.55	3.11E-03	215	218	4			4			N6	N7
AT4G37925	**NdhM**	7.01	5.24E-02	209	217	3			1	2		N7	N8
AT1G15980	**Ndh48/PnsB1**	8.6	1.77E-02	465	461	7	1		4	2		N6	N7
AT1G18730	**NdhB4/PnsB4**	8.8	9.57E-03	182	174	5		1	2	2		N6 and N7	N7
AT5G52100	**CRR1**	4.05	6.78E-03	302	298	5			1	1	3	after N7	N3
Gene in photorespiration pathway													
AT5G35630	**GSL1**	0.08	2.15E-02	430	430	8	3		1	2	1	N7	N7
AT1G32470	**GDC-H**	0.23	7.29E-01	162	166	6+2-aa INS + 1-aa INS				5 + 2-aa INS + 1-aa INS	1	N7	N7
AT4G37930	**SHM**	0.16	4.08E-01	517	517	8	3			5	1	N7	N7
AT1G80380	**GLYK**	0.49	3.61E-01	443	456	8			2	6	1	N7	N7
AT5G04140	**GOGAT**	0.57	1.32E-01	1616	1648	18	4		>=1	>=5	2	N7	N7

Figures

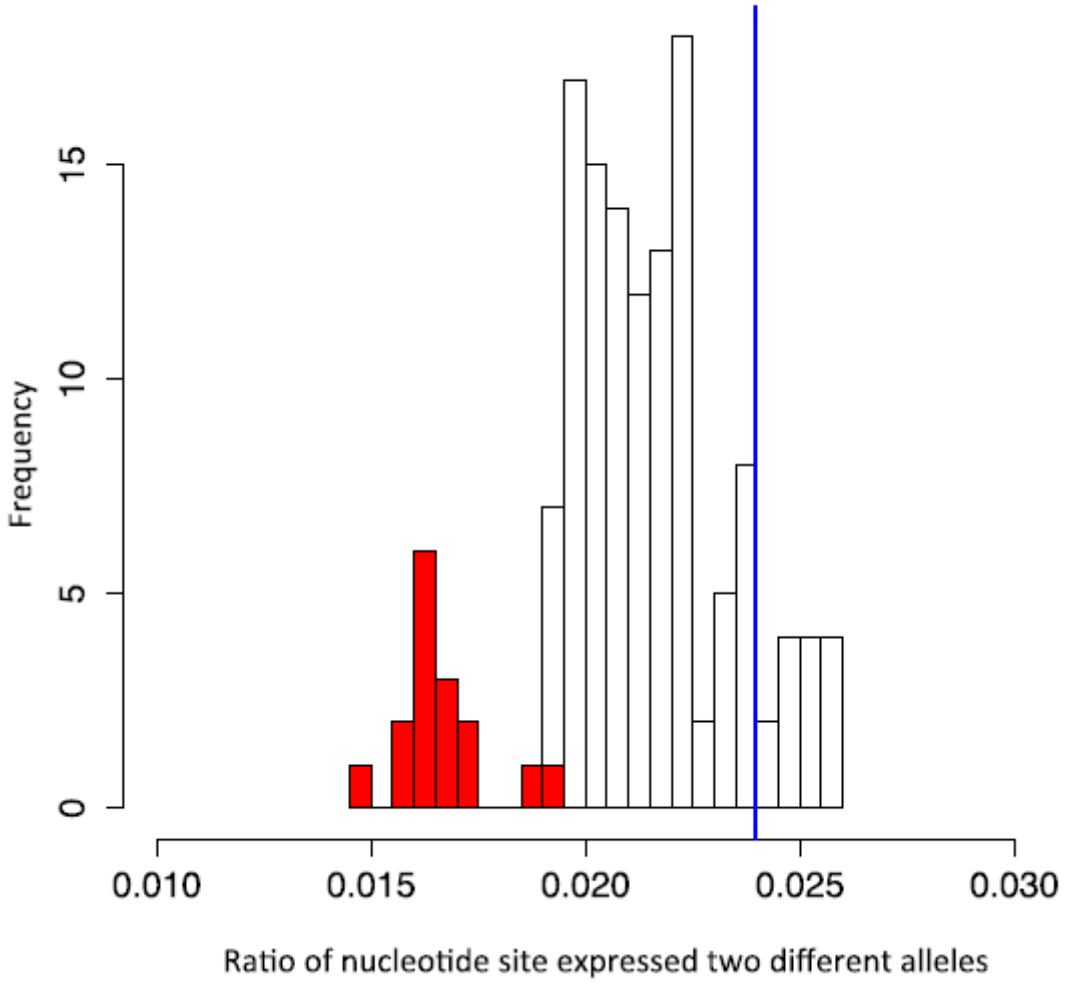

Figure 1

Estimation of the probability of RNA-Seq data from hybrid species The bars show the distribution of ratio of nucleotide sites expressing two different alleles (mixed site). Mixed RNA-Seq samples are generated by pair-wise mixing RNA-Seq data of 16 Flaveria species, which mimics the case of hybridization. The ratio of mixed site in the mixed RNA-Seq samples is showed in grey bars (positive control). The ratio of mixed site of 16 Flaveria species is showed in red bars (real causes). The ratio of hybrid sample F. pringlei*, was represented in blue vertical line.

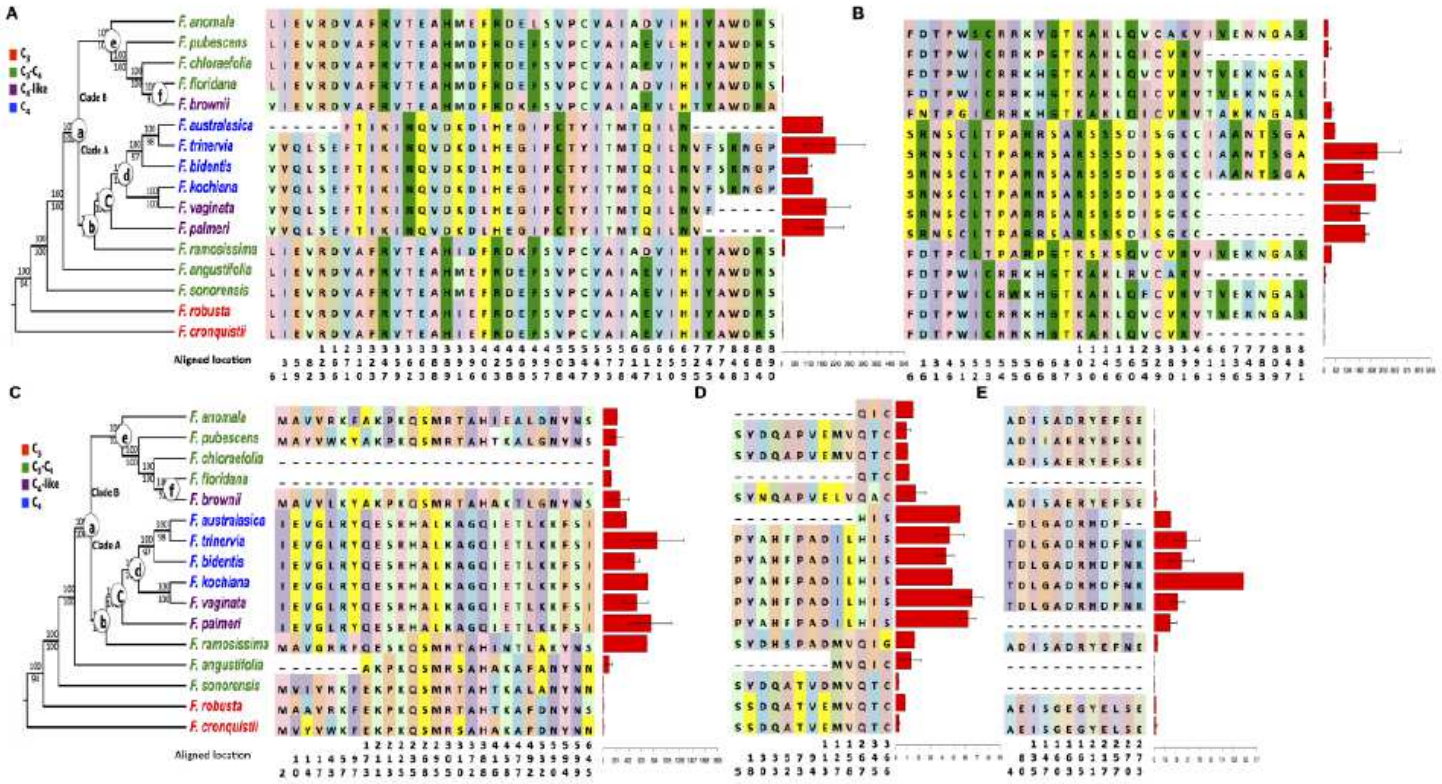

Figure 2

modiscation. Ce pathwaY M Predicted sequences and ,ranscriPt abundances mapped to the Flaverirt phylogeny The predicted amino acid changes and the transcript abundance (FPKM) of the genes [needing the enzymes in CI pathway are shown. Only the amino acid residues predicted to he different between Ca and C. species are superimposed on Me schema of Mover. phylogenetic tree modified from Lyn Mal_ 2015. The colors of amino acid residues have no meaning and are ody for visualization purposes. Numbers below the amino acids indicate the location sites in the multiple sequence alignments. FPKM values with standard errors are shown to the right of Me amino acid changes as red bars. A: phosphosmolp,vate carboxylase (MC); pyruvate orthophosphate dikinase (PPOK); C: NADP-medic enzyme (NADP-3.); pyruvate orthophosphate damase regulatory protem (PPDK-RP); E: phosphoonolpyruvate protem kmase A (PPCKA). Protein sequences from UmprotKP are: F trinervia PEPC, P50694; 7, biderrits WOK. Q59755; F bmw. PPOK., Q39734; and F Irinervia PM., P22221. Sequence alignments are available in Additional file 2

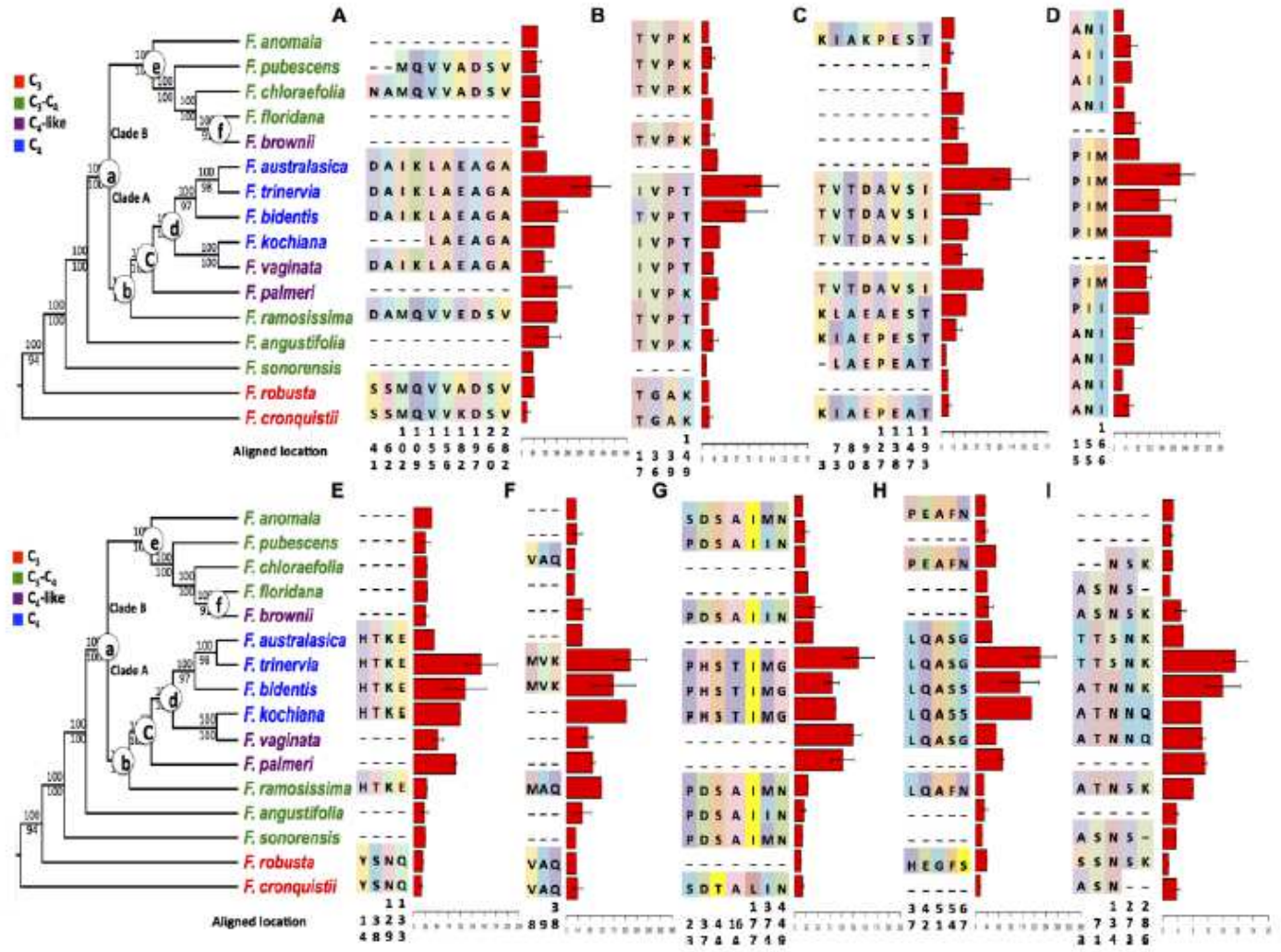

Figure 3

Modifications in the predicted amino acid sequences of proteins involved in cyclic electron transport and transcript abundances of the cognate fransuipfs mapped to the novena phylogeny Changes in predicted amino acid sequence in proteins involved in cyclic electron transport chum and abundances (FPKM) of their cognate transcripts in C., and Cr Flaverza species are shown. Only the amino acid residues predicted to be different between C, and C4 species are superimposed on the schema of Floverm phylogenetic tree modified from Lyu et al., 2015. The marked colors of amino acid residues have no meaning and are only for visualization proposes. Numbers below the amino acids indicate the location site in the multiple sequence alignments. PPKM values with standard errors are represented to the right of the ammo acid changes as red bars. A: protein gradient regulation 5 like protein (PGR5-like); B: NADH dehydrogenase-like (Ndh) L2 submit (Ndh L2); C: D: Ndh16; E: NdhU; F: NdhM; G: Ndh48; H: NdhB4; 1: chlororespiration reduction I (CRR1). The sequence alignments are available in Additional file 2.

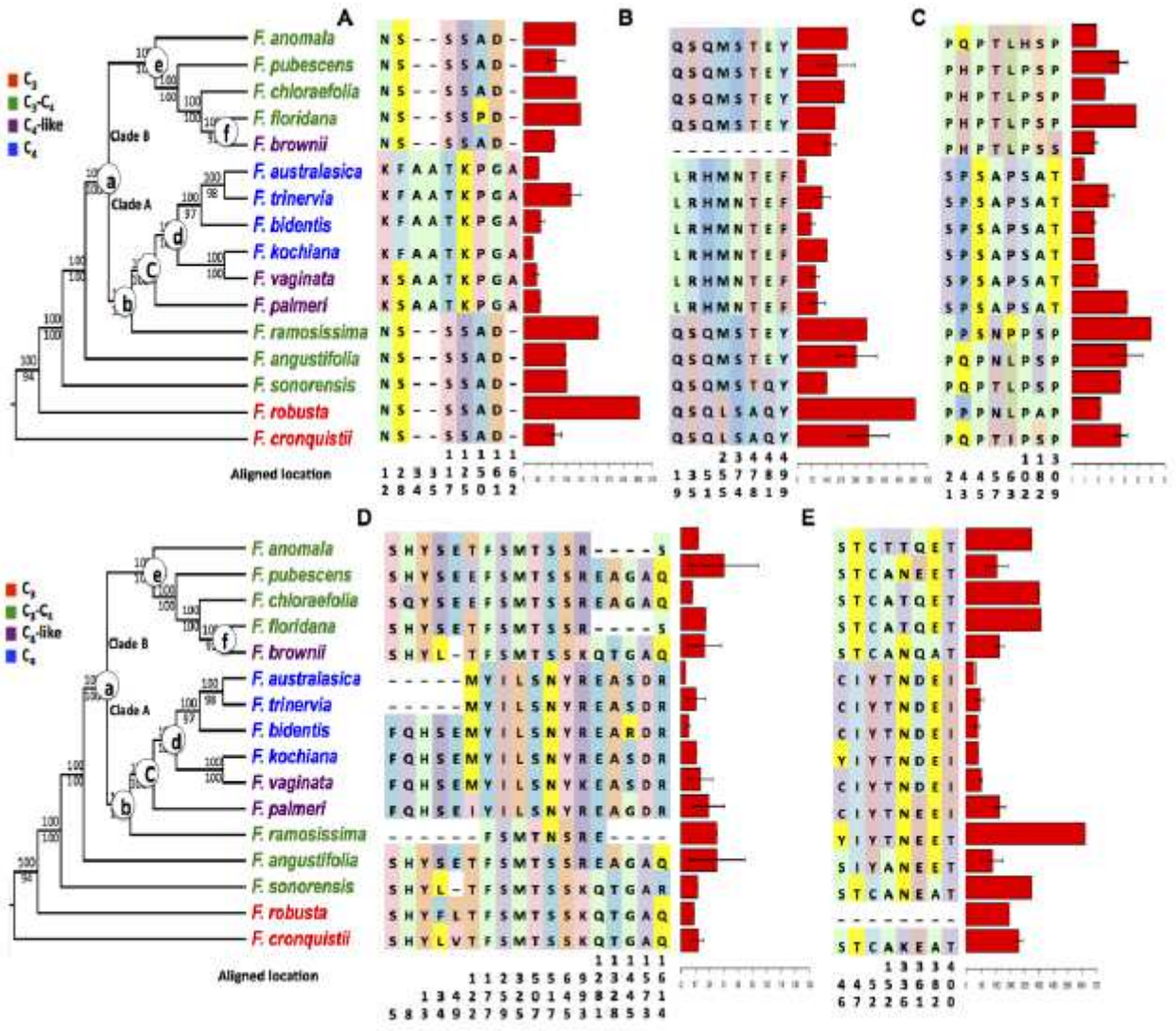

Figure 4

Modifications in photorespiratory protein predicted amino acid sequences and cognate transcript abundances mapped to the Haveria phylogeny The predicted amino acid changes in photorespiratory proteins between Cs and Cs Flaveria species and the transcript abundance (FPKM) of genes encoding the proteins are shown. Only the amino acid residues that are predicted to be different between Cs and Co species are superimposed on the schema of Flaverm phylogenetic tree modified from Lyu et al., 2015. The marked colors of amino acid residues have no meaning and are only for visualization purposes. Numbers below the amino acids indicate the location site in the multiple sequence alignments. FPKM values with standard errors are represented to the right of the amino acid changes as red bars. A, Glutamine synthetase like 1 (GSL1); B: glycine decarboxylase complex H subunit (GDC-H); C: serine hydroxymethyltransferase (SHM); D: glycerate kinase (GLYK); E: glutamine synthetase and glutamine oxoglutarate aminotransferase (GOGAT).

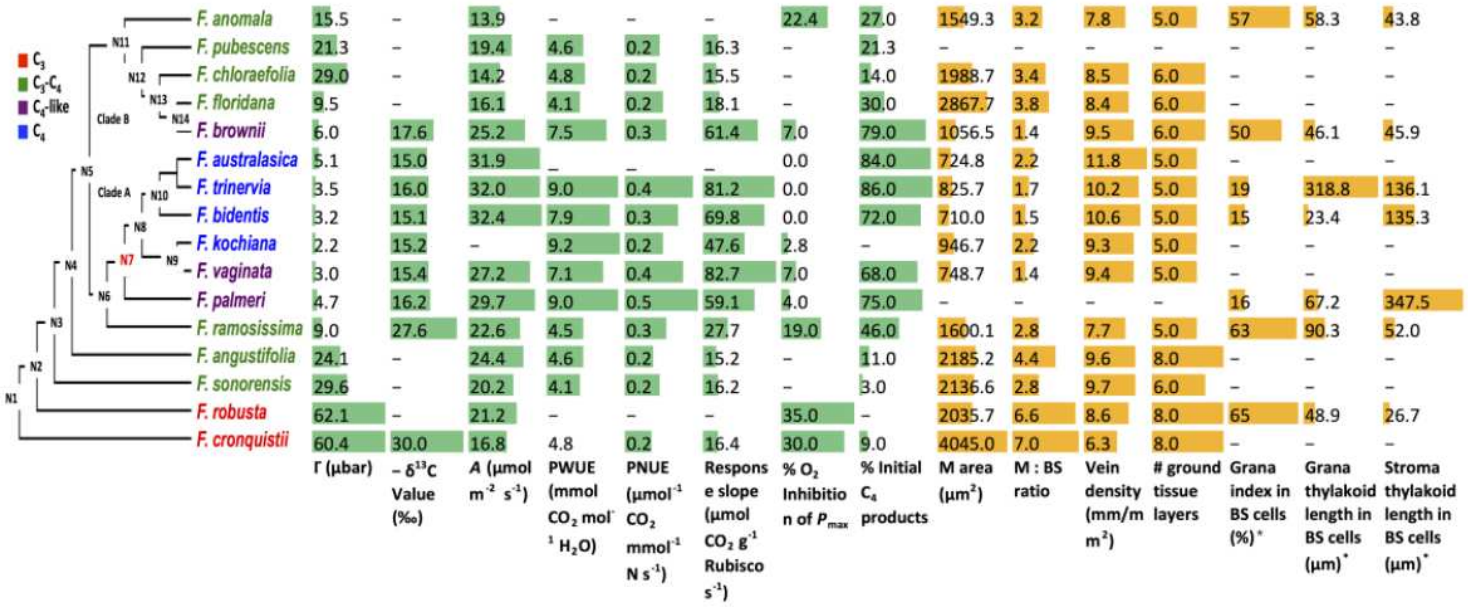

Figure 5

Changes in physiological and anatomical traits mapped onto the Flaveria phylogeny Overall, Ca-related physiological (green and blue bars) and anatomical traits (orange and red bars) showed a step-wise change along the Flaveria phylogenetic tree; however, a number of the traits showed greater more significant changes at certain nodes. *Grans index: total length of grana/total length of thytakoid membrane X 100. (Abbreviations: F: CO2 compensation point; A: CO, assimilation rate; PWUE: instantaneous photosynthetic water use efficiency; PNUE: instantaneous photosynthetic nitrogen use efficiency; response slope: slope of the response of net CO, assimilation rate versus leaf Rubisco content; M: mesophyll; BS: bundle sheath.) Data are from references as given in the Methods.

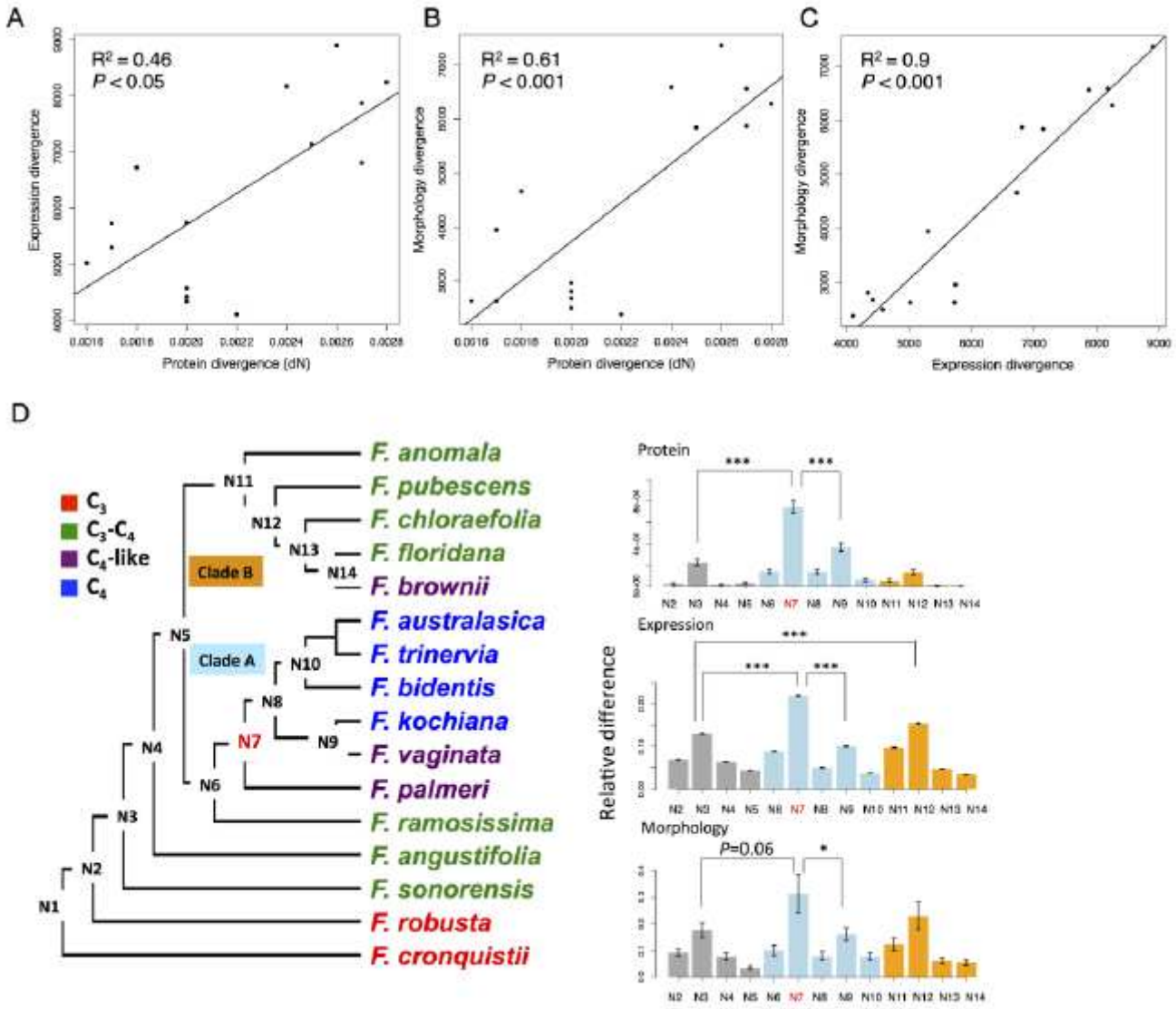

Figure 6

Coordinated evolu,n of protein sequence, gene expression and morphology with an obvious jump change Significant linear conelation between protein divergence, gene expression divergence and morphology divergence were showed in (A-C). Protein divergence was calculated as non-synonymous mutation OM. Expression divergence and morphology divergence were calculated as Euclidean distance based on quantile normalized EPICIVI values and coded morphology values from Mckown eta/. 2005, respectively. All the relative divergences were the divergence between F cmnquislii and other Flavor. species. (D) Shows the relative difference of each ancestral node compared with its earlier ancestral node in protein sequence, gene express, and morphology_ The left panel shows the schema of Mover. phylogenetic tree modified from Lyn et a1. 2015. Each ancestral node eras numbered in the evolutionary sequential order. P values are from One-way ANOVA analysis followed by Mike}, Post Hoc test and adjusted by Barionrin-liochberg cm-rechou The significant levels are:': P<0.05; P.:0.001. The bar colors in grey/blueforange ,resent species from basalklade Aklade B of phyloge,ic nee, resPectivelY

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1supplementarymethodsfigurestables.docx
- Additionalfile2proteinalignment.pdf
- Additionalfile3MeanFPKM.xlsx