The Coordination and Jumps along C₄ Photosynthesis

Evolution in the Genus *Flaveria*

Ming-Ju Amy Lyu^{1,2}, Udo Gowik³, Peter Westhoff³, Yimin Tao², Steve Kelly⁴, Sarah Covshoff⁵, Harmony Clayton⁶, Julian M. Hibberd⁵, Rowan F. Sage⁷, Martha Ludwig⁶, Gane Ka-Shu Wong^{8,9,10}, Xin-Guang Zhu^{1,2§}

1. Institute of Plant Physiology and Ecology, Chinese Academy of Sciences, Shanghai, China

2. Key Laboratory of Computational Biology and National Key Laboratory of Hybrid Rice, CAS-MPG

Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Chinese

Academy of Sciences, Shanghai 200031, China

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

Germany

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

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

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

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

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

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

10. Department of Medicine, University of Alberta, Edmonton AB, T6G 2E1, Canada

§Corresponding author

Corresponding author: Email: zhuxg@sippe.ac.cn, Tel: 86-21-54920486

25 **Abstract**

26 **Background:** C₄ photosynthesis is a remarkable complex trait, elucidations of the
 27 evolutionary trajectory of C₄ photosynthesis from its ancestral C₃ pathway can help us
 28 to better understand the generic principles of complex trait evolution and guide
 29 engineering of C₃ crops for higher yields. We used the genus *Flaveria* that contains C₃,
 30 C₃-C₄, C₄-like and C₄ species as a system to study the evolution of C₄ photosynthesis.

31 **Results:** We mapped transcript abundance, protein sequence, and morphological
 32 features to the phylogenetic tree of the genus *Flaveria*, and calculated the evolutionary
 33 correlation of different features. Besides, we predicted the relative changes of ancestral
 34 nodes of those features to illustrate the key stages during the evolution of C₄
 35 photosynthesis. Gene expression and protein sequence showed consistent modification
 36 pattern along the phylogenetic tree. High correlation coefficients ranging from 0.46 to
 37 0.9 among gene expression, protein sequence and morphology were observed, and the
 38 greatest modification of those different features consistently occurred at the transition
 39 between C₃-C₄ species and C₄-like species.

40 **Conclusions:** Our data shows highly coordinated changes in gene expression, protein
 41 sequence and morphological features. Besides, our results support an obviously
 42 evolutionary jump during the evolution of C₄ metabolism.

43

44 **Key words**

45 C₄ photosynthesis, evolution, coordination, jump, *Flaveria*

46 **Background**

47 Elucidating the evolutionary and developmental processes of complex traits
48 formation is a major focus of current biological and medical research. Most health
49 related issues, including obesity and diabetes, as well as agricultural challenges, such as
50 flowering time control, crop yield improvements, and disease resistance, are related to
51 complex traits [1-3]. Currently, genome-wide association studies are used to study
52 complex traits. Putative genes or molecular markers of importance are then evaluated
53 by a reverse genetics approach to identify those influencing the complex trait. C_4
54 photosynthesis is a complex trait that evolved from C_3 photosynthesis. When compared
55 with C_3 plants, C_4 plants have higher water, nitrogen and light use efficiencies [4].
56 Interestingly, C_4 photosynthesis has evolved independently more than 66 times,
57 representing a remarkable example of convergent evolution [5]. Accordingly, C_4
58 evolution is an ideal system for investigation of the mechanisms of convergent
59 evolution of complex traits.

60 C_4 photosynthesis contains a number of biochemical, cellular and anatomical
61 modifications when compared with the ancestral C_3 photosynthesis [6, 7]. In C_3
62 photosynthesis, CO_2 is fixed by ribulose-1,5-bisphosphate carboxylase/oxygenase
63 (Rubisco), whereas in dual-cell C_4 photosynthesis, CO_2 is initially fixed into a
64 four-carbon organic acid in mesophyll cells (MCs) by phosphoenolpyruvate
65 carboxylase (PEPC) [8]. The resulting four-carbon organic acid then diffuses into the
66 bundle-sheath cells (BSCs) [9], where CO_2 is released and fixed by Rubisco. Hence, C_4

67 photosynthesis requires extra enzymes in CO₂ fixation in addition to those already
 68 functioning in C₃ photosynthesis, including PEPC, NADP-dependent malic enzyme
 69 (NADP-ME), and pyruvate, orthophosphate dikinase (PPDK) [8]. In dual-cell C₄
 70 photosynthesis, CO₂ is concentrated in enlarged BSCs that are surrounded by MCs,
 71 forming the so-called Kranz anatomy [10-12]. Compared with C₃ leaf anatomy, Kranz
 72 anatomy requires a spatial rearrangement of MCs and BSCs, cell size adjustment for
 73 increased numbers of organelles, larger organelles and metabolite transfer between the
 74 two cell types, and a reduction in distance between leaf veins.

75 Much of the current knowledge regarding the evolution of C₄ photosynthesis was
 76 gained through comparative physiological and anatomical studies using genera that
 77 have not only C₃ and C₄ species, but also species performing intermediate types of
 78 photosynthesis [7, 13]. Among these, *Flaveria* has been promoted as a model for C₄
 79 evolution studies [14], and the evolution of C₄-related morphological, anatomical and
 80 physiological features has been well studied in this genus over the last 40 years [14-17].
 81 Though the molecular evolution of several key C₄ enzymes were reported in this genus
 82 [18-20], however, the molecular evolution of C₄ related features is large unknown.
 83 Besides, the evolutionary relationship of the C₄ related molecular features and
 84 morphology features is not clear so far. In this study, we combined transcriptome data
 85 and published morphology data, together with the most recent phylogenetic tree of the
 86 genus *Flaveria* [21], to systematically investigate the key molecular events and
 87 evolutionary paths during the C₄ evolution. Our results revealed that though many of

the changes related to C₄ photosynthesis occurred gradually, there are strong coordination and evolutionary jumps 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 Archive (SRA) of the National Center for Biotechnology Information (NCBI) (Table S1). The 16 species represented two C₃ species, seven C₃-C₄ intermediate species, three C₄-like species and four C₄ species [22, 23] (Table S1). On average, 42,132 contigs (from 30,968 to 48,969) with lengths of no less than 300 bp were assembled for each of the 16 species. The N50 of these contigs ranged from 658 to 1208 (Table S2). The 16 species had a similar contig length distribution, with a peak at around 360 bp (Fig. S1). Since *Flaveria* is a eudicot genus, we used *Arabidopsis thaliana* (*Arabidopsis*) as reference to annotate *Flaveria* transcripts. On average, 58.91% of *Flaveria* contigs had orthologous genes in *Arabidopsis*.

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 correlations in overall transcript abundance in leaf samples from the same species than those of different organs from the same species, regardless of source (Fig. S2). Specifically, leaves from different developmental stages or from different labs are more closely correlated than leaf samples from different

108 species, or than mean values of pair-wise correlations across all 27 leaf samples (T-test,
109 $P < 0.05$) (Fig. S3). Therefore, the mean FPKM value from multiple leaf samples was
110 assigned as the final transcript abundance of the leaf for each species. Our
111 quantification showed that, in *F. bidentis* (C_4), transcripts from genes encoding
112 C_4 -associated proteins were more abundant in leaf than in root and stem tissues (Fig.
113 S4), which is consistent with earlier reports [24, 25]. In contrast, in the C_3 species *F.*
114 *robusta*, the difference in transcript abundance of orthologous genes between leaf and
115 root/stem tissues was much less. Our data showed that NADP-ME is the dominant C_4
116 pathway in *Flaveria* species (Fig. S5). We also proved that all species used in this study
117 are from natural evolution and can be used for this evolutionary study (Fig. S6, see
118 Supplementary Results). As a result, 13,081 Arabidopsis orthologues were detected in
119 at least one of the 16 *Flaveria* species, and 12,215 genes were kept with the maximum
120 FPKM in 16 species no less than 1 FPKM.

121 **C_4 related genes: genes showed difference in gene expression and protein** 122 **sequence between C_3 species and C_4 species**

123 We first identified C_4 related genes, which were defined as genes show difference
124 in both gene expression and protein sequence between C_3 and C_4 species. We first
125 calculated the differentially expressed (DE) genes between C_3 and C_4 species, which
126 resulted in 2,306 DE genes. We next investigated transcriptome-wide amino acid
127 changes predicted from orthologues of C_3 and C_4 *Flaveria* species using the process
128 shown in Fig. S7 (Supplementary Methods). To estimate the accuracy of the predicted

129 peptide sequences from our data, we conducted a comparative study of protein
130 sequences from UniProtKB (<http://www.uniprot.org>) with those from our data, we
131 found that our predicted peptide sequence is as good as, if not better than, the sequence
132 from UniProtKB in terms of accuracy (details see Supplementary Results and Table S3).
133 As a result, we obtained 1,018 genes encoding at least one amino acid change between
134 C₃ and C₄ *Flaveria* species. 205 out of these 1,018 genes also showed significantly
135 differentially expression ($P < 0.01$) between C₃ and C₄ species, which was termed as C₄
136 related genes, 113 and 92 showed ascending and descending transcript abundance in C₄
137 species relative to C₃ *Flaveria* species, respectively (Fig. S8).

138 We then investigated the degree of overlap of the 205 C₄ genes from *Flaveria* with
139 genes known or having the potential to be related to C₄ photosynthesis or C₄ evolution
140 in different species, including *Arabidopsis* [26], (Fig. S9), *Gynandropsis gynandra* [27]
141 (Fig. S10), *Setaria viridis* [28, 29] (Fig. S11) and *Zea mays* (maize) [30] (Fig. S11 and
142 Fig. S12). Result shows that the 205 genes are significantly enriched in those genes that
143 are potentially related to C₄ photosynthesis or C₄ evolution ($P < 0.05$, “BH” adjusted)
144 (details see the Supplementary Results).

145 **The C₄ related genes showed coordinated and abrupt change along the C₄** 146 **evolutionary pathway in the genus *Flaveria***

147 The 205 genes are significantly enriched in several gene ontology (GO) terms
148 including photosynthesis, photorespiration, photosynthetic light reaction,
149 photosynthesis electron transport in photosystem I (PSI), photosynthesis light

150 harvesting in PSI, chloroplast, organic anion transport and oxidation-reduction ($P < 0.05$,
151 Fisher's exact test, "BH" adjusted; Table S4). In the following sections, we
152 systematically discuss these genes and their changes during C_4 evolution in *Flaveria*
153 with regard to gene expression and predicted protein sequences. We first focus on genes
154 encoding proteins associated with C_4 photosynthesis, then on genes related to the
155 enriched GO terms (Table S4), which satisfy the following two criteria: (1) more than
156 two predicted amino acid differences between C_3 and C_4 species, and (2) fully
157 assembled predicted protein sequences from species belonging to all four
158 photosynthetic types: C_3 , C_3 - C_4 , C_4 -like and C_4 . In general, these genes were classified
159 into six categories according to the probable function of their cognate proteins, *i.e.*, the
160 C_4 pathway, photorespiratory pathway, electron transport chain, membrane transport,
161 photosynthetic membrane, and oxidation-reduction.

162 *Genes encoding proteins associated with the C_4 pathway*

163 Nine genes encoding proteins associated with the C_4 pathway were identified,
164 including those encoding three C_4 cycle enzymes, PEPC, PPDK and NADP-ME, two
165 regulatory proteins, PPDK regulatory protein (PPDK-RP) and PEPC protein kinase A
166 (PPCKA), two aminotransferases, Alanine aminotransferase (AlaAT) and aspartate
167 aminotransferase 5 (AspAT5), and two transporters, BASS2 and sodium: hydrogen
168 (Na^+/H^+) antiporter 1 (NHD1) (Table 1). In terms of protein sequence, the major
169 predicted amino acid changes in C_4 species occurred at N7 for all of the nine genes
170 (Figs. 1, 2, Figs. S13, Table 1). For example, PEPC in the C_4 *Flaveria* species had 41
171 predicted amino acid changes compared with those in the C_3 species, which were

mapped onto the *Flaveria* phylogeny determined by Lyu *et al.* [21]. One of the predicted changes occurred at N6 (D396 in C₄ species, hereafter D396), and 34 occurred at N7 (Fig. 1A). The six other predicted amino acid changes occurred at N7 or after N7, although the incomplete assembly of PEPC transcripts from *F. palmeri* and *F. vaginata* did not allow resolution of the predicted amino acid sequences. These results suggest an evolutionary jump in the protein sequence at N7 for C₄ enzymes.

In terms of gene expression, all the nine genes showed higher transcript abundance in C₄ species than in C₃ species and a comparable level in C₄-like and C₄ 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, C₄ species showed a 7.6-fold to 123.6-fold of FPKM values compared with C₃ 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. 2A and Fig. S13A). Our results hence suggested that the genes encoding proteins associated with C₄ pathway showed highly coordinated modification patterns in protein sequence and gene expression at N3, N6 and N7 during the evolutionary pathway of C₄ photosynthesis; while the majority of the predicted amino acid changes occurs at the N7.

Genes encoding proteins in the photorespiratory pathway

Transcripts encoding five proteins involved in photorespiration were identified in

our comparative analyses: glycine decarboxylase complex (GDC) H subunit (GDC-H), serine hydroxymethyltransferase (SHM), glycerate kinase (GLYK), glutamine synthetase and glutamine oxoglutarate aminotransferase (GOGAT), and glutamine synthetase-like 1 (GSL1) (Figs. 3, Table 1). In general, the predicted amino acid substitution patterns of these five proteins were similar to those observed in the above described proteins in C₄ pathways, with the major predicted amino acid changes in C₄ species occurring at N7 (Figs 3, Table 1), *e.g.*, 16 of 18 in GOGAT occurred at N7 (Fig 3D). Generally, proteins in the photorespiratory pathway showed fewer predicted amino acid changes than those in the C₄ pathway.

The abundance of transcripts encoding the five photorespiratory enzymes examined was comparable in C₃ and C₃-C₄ species, and higher than in the C₄ species (Figs. 3 A–E). Interestingly, we found a consistent pattern in transcript abundance that paralleled the trajectory of C₄ evolution, with the highest, or the second highest FPKMs found in the C₃-C₄ species *F. ramosissima* and the lowest observed in the C₄ species. For example, the FPKM values for GOGAT were 30.73 in *F. robusta* (C₃), 83.97 in *F. ramosissima* (C₃-C₄), 56.95 in *F. palmeri* (C₄-like) and 27.27 in the C₄ species *F. kochiana* in clade A. In clade B, the values changed from 30.73 in the C₃ *F. robusta* to 35.35 in *F. anomala* (C₃-C₄), 92.01 in *F. pubescens* (C₃-C₄), and 26.27 in *F. brownii* (C₄-like). Moreover, transcripts encoding photorespiratory enzymes exhibited at least a 1.5-fold difference between C₄-like species and C₄ species in clade A. Specifically, when compared with *F. kochiana* (C₄), *F. vaginata* (C₄-like) showed a 1.55-fold increase for GOGAT (27.27 to 42.23), a 1.54-fold increase for GSL1 (86.99 to 133.59),

216 and a 1.51-fold increase for GDC-H (37.06 to 56.06). *F. palmeri* displayed an even
217 larger fold-change relative to *F. kochiana* than *F. vaginata* (Figs. 3 A-E). Hence, when
218 compared with genes encoding C₄ pathway proteins, those encoding photorespiratory
219 proteins showed larger differences between C₄-like and C₄ species in clade A in terms
220 of gene transcript abundance and cognate protein sequence. The greatest reduction of
221 FPKM was observed uniformly at N7 (Figs. 3, Table 1). Thus, this suggested that the
222 genes encoded proteins associated with photorespiratory pathway also showed
223 coordinated changing pattern in protein sequence and gene expression during the
224 evolutionary pathway of C₄ photosynthesis, again with the largest number of changes
225 occurring at N7.

226 *Genes encoding proteins involved in the electron transport chain*

227 We identified genes encoding 12 proteins that function in the photosynthetic
228 electron transfer chain, including nine related to cyclic electron transport (CET) (Fig. 4),
229 PSI light harvesting complex gene 5 (Lhca5), post-illumina chlorophyll fluorescence
230 increase (PIFI) and cytochrome b561 (Cytb561). Transcripts encoding all 12 of the
231 proteins showed higher abundances in C₄ species than C₃ species (Figs. 4, Table 1). The
232 genes encoding proteins involved in CET showed the biggest changes at different nodes
233 instead of at a single node, *e.g.*, the major changes of PGR5-like in FPKM and in
234 predicted protein sequences occurred at N6 and N7, respectively (Fig. 4A). Transcripts
235 encoding NdhL2 showed the biggest increase in abundance at N7, and two of four
236 predicted amino acid changes occurred before N5 (Fig. 4B); the major changes of

237 Lhca5 were observed at N5 (Fig. S14A). This suggested that the variation of CET
238 might have contributed to split of clade A and B in the *Flaveria* genus.

239 *Genes encoding proteins associated with photosynthesis, transport, and*
240 *oxidation-reduction*

241 Two genes labelled in GO term as photosynthesis, namely, Rubisco activase (RAC)
242 and orthologue of AT5G12470 were identified. RAC showed the greatest decrease in
243 FPKM at N7, while nine out of eleven predicted modifications in its sequence were
244 acquired at N6 (Fig. S15A, Table 1). The orthologue of AT5G12470 in *Flaveria*,
245 (hereafter *Flaveria*-AT5G12470), a chloroplast envelope protein [31] potentially
246 involved in responses to nitrate levels [32], showed the greatest changes in transcript
247 abundance and predicted protein sequences at N7 (Fig. S15B, Table 1).

248 Transcripts encoding two transport proteins also displayed changes in FPKM and
249 predicted amino acid sequence in C₄ species and C₃ *Flaveria* species, namely,
250 multidrug and toxic compound extrusion (MATE) protein and amino acid permease 6
251 (AAP6) (Fig. S16A and B). AAP6 was reported to be a high affinity neutral amino acid
252 transporter primarily expressed in sink tissue and xylem parenchyma cells, and
253 potentially responsible for taking up amino acids from the xylem and delivering them
254 to the phloem [33-35]. The modifications in expression levels were observed at N9 for
255 MATE and N8 for AAP6, and the greatest modifications in predicted sequence were
256 observed at N6 and N7, respectively (Fig. S16, Table 1).

257 Transcripts encoding two proteins playing roles in oxidation-reduction showed

changes between C₄ and C₃ species of *Flaveria*, namely, glutathione reductase (GR) and sorbitol dehydrogenase (SorDH). GR showed the major enhancement of transcript abundance at N3 and predicted amino acid changes at N6 and N7 (Fig. S17A, Table 1). Similarly, the major reduction of transcript abundance of and predicted amino acid changes of SorDH showed at N7 (Fig. S17B). These results are consistent with studies suggesting a pivotal role of redox status in the expression of genes encoding photosynthetic proteins [36, 37].

Physiological and anatomical characteristics related to C₄ photosynthesis show coordinated changes along the C₄ evolutionary pathway in *Flaveria*

To investigate whether C₄-related physiological characteristics also underwent coordinated changes during the evolution of C₄ photosynthesis in *Flaveria*, physiological characteristics taken from the literature [23, 38, 39] 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 [14, 23, 38, 39] (Fig. 5); however, coordinated and abrupt changes were observed for a number of features. A major change in CO₂ compensation point (Γ) in *Flaveria* was first seen at N3, where the most ancestral C₃-C₄ species, *F. sonorensis*, was emerged which showed a decrease in Γ from 62.1 μ bar of its closest C₃ relative *F. robusta* to 29.6 μ bar (Fig. 5). The greatest changes in Γ in clade A occurred at N6, which showed a decrease in Γ from 24.1 μ bar in *F. angustifolia* (C₃-C₄) to 9.0 μ bar in *F. ramosissima* (C₃-C₄), followed by N7, where a decrease in Γ from 9.0 μ bar in *F. ramosissima* (C₃-C₄) to 4.7 μ bar in *F. palmeri* (C₄-like)

279 was seen. The greatest decrease of Γ in clade B was observed between the two C₃-C₄
 280 species, *F. floridana* and *F. chloraefolia* (C₃-C₄), where there was a decrease from 29
 281 μ bar to 9.5 μ bar (Fig. 5). For photosynthetic water using efficiency (PWUE),
 282 photosynthetic nitrogen using efficiency (PNUE) and the slope of the response of the
 283 net CO₂ assimilation rate (*A*) versus Rubisco, the biggest changes occurred at N7 with
 284 increases of around 2-fold. In contrast, the percentage of ¹⁴C fixed into four carbon
 285 acids showed no clear trend along the phylogenetic tree, although 3.91-fold and
 286 1.76-fold increases were seen at N6 and N7, respectively. Interestingly, changes in all
 287 of these traits uniformly occurred at *F. brownii* in clade B, the only C₄-like species
 288 within this clade. Consequently, those data suggest that although there were gradual
 289 changes in physiological features along the C₃, C₃-C₄, C₄-like and C₄ trajectory, there
 290 are apparent jumps at N3, N6 and N7 in these physiological traits along the *Flaveria*
 291 phylogeny (Fig. 5).

292 Anatomical traits [14] were mapped onto the *Flaveria* phylogeny to investigate
 293 how these features were modified along the evolution of C₄ (Fig. 5). For both the area
 294 of MCs and the ratio of the area of MCs to that of BSCs (M: BS), the greatest
 295 modifications along the phylogeny were found between *F. brownii* (C₄-like) and *F.*
 296 *floridana* (C₃-C₄), with a similar degree of change for both characteristics (2.7-fold, Fig.
 297 5). Anatomical data for *F. palmeri* (C₄-like) in clade A are not available; however, large
 298 differences in anatomical features were found between the C₄-like *F. vaginata* and
 299 C₃-C₄ *F. ramosissima* [14]. The modification of M area first occurred at N2 which
 300 showed a 1.9-fold difference (compare 4045 m² in *F. robusta* to 2035.7 m² in *F.*

301 *cronquistii*) followed by a 2.1-fold of difference between *F. ramosissima* and *F.*
302 *vaginata* (compare 1600.1 m² in *F. ramosissima* and 748.7 m² in *F. vaginata*). The
303 major modification of the ratio of M and BS occurred at N3 with a 2.4-fold difference
304 (compare 6.6 m² in *F. robusta* and 2.8 m² in *F. sonorensis*) and N6 with a 1.6-fold
305 difference (compare 4.4 m² in *F. angustifolia* and 2.8 m² in *F. ramosissima*) and
306 between *F. ramosissima* and *F. vaginata* (1.4 m²) with a 2-fold difference. Therefore,
307 similar to the evolutionary pattern of physiological features, large changes in
308 anatomical features also emerged at N3, N6 and the transition between C₃-C₄ species
309 and C₄-like species. Interestingly, the ultrastructure of BSCs chloroplasts showed an
310 abrupt change at N7, with a dramatic decrease in grana thylakoid length and an increase
311 in stroma thylakoid length, whereas these features were comparable in the species at the
312 base of tree and in clade B [40]. These findings are consistent with the observation that
313 the abundance of transcripts encoding proteins involved in CET increased more in
314 clade A species than in clade B species, as described above. These results imply that
315 CET only increased in clade A species, which may be a key factor determining the
316 possibility of forming C₄ species from the C₃-C₄ intermediate species.

317 **Coordinated change of protein sequence, gene expression and morphology**
318 **with an evolutionary jump at the transition between C₃-C₄ and C₄-like**
319 **species along the species evolution**

320 Our above analysis showed that C₄ related features showed coordinated changes
321 with an obvious abrupt change at N7. Next, we asked whether species evolution also

322 showed evolutionary coordination and jump(s) along the species evolution in protein
 323 sequence, gene expression and morphology. To answer this question, we calculated the
 324 divergence matrices for protein sequence, gene expression, and morphological features
 325 between *F. cronquistii* (at the most basal place in the *Flaveria* phylogenetic tree) and
 326 other *Flaveria* species. The protein divergence was calculated as the rate of
 327 non-synonymous substitutions (dN) of all the genes that were used to construct the
 328 *Flaveria* phylogenetic tree from [21]; and expression divergence as Euclidean distance
 329 of total expressed genes (see Methods); and morphology divergence as Euclidean
 330 distance using previously coded morphology value from [41], which includes 30 types
 331 of morphology traits, such as life history, leaf shape, head types and so on. Our data
 332 showed a high linear correlation between the protein divergence, gene expression
 333 divergence and morphology divergence, in particular between gene expression
 334 divergence and morphology divergence ($R^2=0.9$) (Figs. 6A-C), suggesting a
 335 coordinated evolution of protein sequence, gene expression and morphology in species
 336 evolution.

337 Next, we predicted the protein sequence, transcript abundance and coded
 338 morphology value of ancestral nodes, which were then used to calculate the relative
 339 change of the three parameters at each node (see Methods). Surprisingly, protein
 340 sequence and gene expression showed significantly more changes at N7 than changes
 341 at other nodes ($P<0.001$, Tukey's test, "BH" adjusted , the same as following), and the
 342 morphology showed the most changes at N7 with a marginal significant *P* value
 343 ($P=0.06$), implying a evolutionary coordination and jump also occurred in species

344 evolution.

345 Consider that N7 is the most recent common ancestor of C₄ and C₄-like species in
 346 clade A, where C₄-like and C₄ species are comparable with respect to the C₄-ness (Figs.
 347 6), it may be possible that the C₄ photosynthesis accelerate the evolution of species. We
 348 then investigate how much total species variances can be explained by C₄ related genes.
 349 Principle component analysis (PCA) showed that species derived from N7 were
 350 distinguished from other species (Figs. 7), which is consistent with the evolutionary
 351 jump at N7. The first component of the 205 C₄ related genes account for 38% of total
 352 variance (Fig. 7C), more than the dataset of genes that expressed in all species (8004
 353 genes), which account for 32% of total variance (Fig. 7A), and the DE genes account
 354 for 27% of total variance (Fig. 7B). Moreover, the 205 C₄ related genes showed same
 355 evolutionary pattern with the total expressed genes and DE genes, which had the
 356 biggest changes at N7 (Figs. 7) ($P < 0.001$). This raises the possibility that the evolution
 357 of species in the *Flaveria* genus might be mainly driven by the evolution of C₄
 358 photosynthesis. This is not surprising considering the generally higher light, nitrogen
 359 and water use efficiency in C₄ photosynthesis as compared with C₃ photosynthesis. It is
 360 highly likely that many parameters related to growth, development and responses to
 361 environments differ between species with these two different photosynthetic pathways.

362 Discussion

363 Evolutionary coordination of different features implies a purifying 364 selection towards a functional C₄ metabolism

365 Compared with C₃ photosynthesis, C₄ photosynthesis acquired many new features
366 in gene expression, protein sequence, morphology and physiology (Figs. 1-6) [42]. We
367 interpret these coordinated changes as a result of a strong purifying selection at this step.
368 This is because though C₄ photosynthesis can gain higher photosynthetic energy
369 conversion efficiency, highly specialized leaf and cellular anatomical features and
370 biochemical properties of the involved enzymes are required. For example, increased
371 cell wall thickness at the bundle sheath cell and decreased sensitivity of PEPC to malate
372 inhibition are needed for C₄ plants to gain higher photosynthetic rates [43, 44].
373 Furthermore, to gain higher photosynthetic efficiency in C₄ plants, the ratio of the
374 quantities of Rubisco content in BSCs and MCs is also critical [45]. In theory, if the C₄
375 decarboxylation is in place occurs before all other accompanying features, leaves will
376 experience high leakage, *i.e.*, costing ATP for a futile cycle without benefit to CO₂
377 fixation. This will inevitably lead to lower quantum yield and a potential driving force
378 for purifying selection. Further evidence for possible purifying selection comes from
379 the observation that genes with cell-specific expression, such as PEPC, PPKK, and
380 NADP-ME, displayed more changes in their predicted protein sequences than
381 ubiquitously expressed genes, such as NDH components (Table 1, Additional file 3).
382 This is because, as discussed earlier, the redox environments between BSCs and MCs

might have changed dramatically during the completion of the C₄ cycle, with one of the most likely change being having a more acidic environment due to increased production of Oxaloacetic acid (OAA) and malate. Under such conditions, it is required for enzymes to alter their amino acid sequences to adapt to the new cellular environments. The concurrent changes between gene expression divergence and protein sequence divergence has also been demonstrated previously in animals [46, 47], which has been similarly proposed to reflect negative selection for the involved genes [47].

Evolutionary jumps along the C₄ evolution in the *Flaveria* genus

Among the nodes leading to the C₄ emergence in clade A, the N7 shows the biggest change in protein sequence, gene expression and morphology in both C₄ specific features and also general features (Table 1, Figs. 1-7). There are also apparent changes in these features at N3 and N6. These three nodes reflect three critical stages during the emergence of C₄ metabolism. First, at N3, there was a large degree of changes in gene expression, protein sequence and morphology (Figs. 1 and 2). One of the most important events during this phase is the re-location of GDC from MSCs to BSCs based on earlier western blot data [13, 48]. Here we found that SHM showed decreased expression while most of other photorespiratory related enzymes showed little changes (Figs. 4). Similarly, at this step, the majority of the C₄ related genes showed little changes (Figs. 1 and 2). In contrast, global survey of the gene expression, protein sequence and morphology changes suggest that there is large number of

changes at this step compared to earlier C₃ species (Figs. 6 and 7), and there is also great decrease of CO₂ compensation point at this stage (Fig. 5).

N6 is the stage where we found the third largest degree of changes occurred in C₄ related features. At this stage, we observed large increase in transcripts coding for nearly all enzymes involved in nitrogen rebalancing (Figs. 1-3), photorespiration related transcripts, and concurrent increase in transcripts encoding the other remaining C₄ cycle-related enzymes, and a dramatic increase in the percentage of ¹⁴C incorporated into the four carbon acids occurred (Fig. 5). The increase in transcript abundance in photorespiratory genes might be related to the optimization of C₂ cycle to decrease CO₂ concentrating point, which can increase fitness of plants under conditions favoring photorespiration [49]. The dramatic increase in enzymes related to nitrogen rebalancing, i.e. PEPC, NADP-ME, PEPCKA etc, is consistent with the notation that C₄ cycle might be evolved as a result of rebalancing nitrogen metabolism after GDC moving from MC to BSC [17]. The fact that there is little change in the $\delta^{13}\text{C}$ in the C₃-C₄ intermediate as compared to that of C₃ species suggests that the contribution of CO₂ fixation following C₄ pathway is relatively minor, i.e., less than 15% estimated based on an $\delta^{13}\text{C}$ value of -27.6 in *F. ramosissima* (Fig. 5), again supporting the initial role of increased C₄ enzymes is not for enhancing CO₂ fixation. It is worth pointing out here that the measured initial carbon fixation in the form of C₄ compound was 46% (Fig. 5), higher than those estimated based on the $\delta^{13}\text{C}$ value. This is possibly because though malate releases CO₂ into BSCs as a result of the nitrogen rebalancing pathway, most of the CO₂ was not fixed by Rubisco, either due to lack of sufficient Rubisco activity in BSCs or

426 due to lack of required low BSCs cell wall permeability to maintaining high CO₂
427 concentration in BSCs *etc.*

428 N7 witnesses abrupt changes for both the gene expression and proteins sequence
429 and morphology (Figs. 1-5, Fig. 6 D). The majority of the C₄ related genes showed the
430 most modification in gene expression and protein sequence at N7, especially for genes
431 in C₄ cycle and photorespiratory pathway. Moreover, N7, where C₄-like species (clade
432 A) appear, represents a dramatic shift of CO₂ fixation from being dominated by a C₂
433 concentrating mechanism to being dominated by a C₄ concentrating mechanism. Based
434 on the $\delta^{13}\text{C}$ value in *F. palmeri*, the fixation through the C₄ concentrating mechanism is
435 up to 93%, which is consistent with the measured proportion of initial carbon fixation
436 in the form of C₄ compound (Fig. 5), suggesting at this step, the released CO₂ in the
437 BSCs can be largely fixed by Rubisco. Whereas, the transition between C₄-like to C₄
438 process is an evolutionarily "down-hill" process and most of optimization occurred
439 through fine-tuning expression abundance (Figs. 1- 4).

440 **Conclusions**

441 Combining transcript abundance, protein sequence, morphology features, here we
442 systematically evaluate the molecular evolutionary trajectory of C₄ photosynthesis in
443 the genus *Flaveria*, in particular the clade A of the genus *Flaveria*. We found a clear
444 evolutionary coordination of different features. Our data also support evolutionary
445 jumps during the evolution of C₄ species, which reflect three major steps during the
446 emergence of C₄ metabolism, including the pre-adaptation step where GDC moved

447 from mesophyll cell to bundle sheath cell (N3), formation of C₂ nitrogen re-balancing
448 pathway and concurrent formation of a C₄ pathway (N6), and dominance of C₄
449 metabolic pathway (N7). The modification at N7 shows the biggest jump during the
450 emergence of C₄ metabolism.

451 **Methods**

452 **Data retrieval**

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

457 CO₂ compensation points (Γ) (except for *F. kochiana*), $\delta^{13}\text{C}$ (except for
458 *F. kochiana*), %O₂ inhibition of P_{max} (except *F. kochiana*), and CO₂ assimilation rates
459 were from [23]. Γ , $\delta^{13}\text{C}$ and %O₂ inhibition of *F. kochiana* were from [38]. Data for %
460 initial C₄ products in total fixed carbon were from [50]. Data for PWUE, PNUE, and
461 net CO₂ assimilation rate (A) versus Rubisco content were from [38]. Data for M area,
462 M:BS ratio, vein density and number of ground tissue layers were from [14]. The
463 values of M area, M:BS ratio and vein density were measured from figures in McKown
464 and Dengler [14] with GetData (<http://www.getdata-graph-digitizer.com>). Mean values
465 from five measurements were used. Ultrastructural data of BS cell chloroplasts were
466 from Nakamura *et al.* [40].

467 **Transcriptome assembly and quantification**

468 Transcripts of *Flaveria* species generated with Illumina sequencing were
469 assembled using Trinity (version 2.02) [51] with default parameters (Table S1). Contigs
470 of four *Flaveria* species from 454 sequencing data were assembled using CAP3 [52]
471 with default parameters. In all cases, only contigs of at least 300 bp in length were
472 saved. Transcript abundances of 31 *Flaveria* samples were analyzed by mapping
473 Illumina short reads to assembled contigs of corresponding species and then
474 normalized to the fragment per kilobase of transcript per million mapped reads (FPKM)
475 using the RSEM package (version 1.2.10) [53]. Functional annotations of *Flaveria*
476 transcripts were determined by searching for the best hit in the coding sequence (CDS)
477 dataset of *Arabidopsis thaliana* (Arabidopsis) in TAIR 10 (<http://www.arabidopsis.org>)
478 by using BLAST in protein space with E-value threshold 0.001. If multiple contigs
479 shared the same best hit in CDS reference of Arabidopsis, then the sum FPKM of those
480 contigs was assigned to the FPKM value of the gene in *Flaveria*. To make the FPKM
481 comparable across different samples, we normalized the FPKM value by a scaling
482 strategy as used by Brawand *et al.* [54]. Specifically, among the transcripts with FPKM
483 values ranking in 20%-80% region in each sample, we identified the 1000 genes that
484 had the most-conserved ranks among 29 leaf samples, which were then used as an
485 internal reference, and the transcript of each sample was normalized according to the
486 mean value of these 1000 genes in the sample. We then multiplied all the FPKM values
487 in all samples by the mean value of 1000 genes in the 29 leaf samples (Fig. S2). Genes

488 showing differential expression were identified by applying dexs (version 1.2.2) [55]
489 in R, with a P value less than 0.05.

490 **Protein divergence, gene expression divergence and morphology**

491 **divergence**

492 Pair-wise protein divergence (dN) was calculated by applying codeml program in
493 PAML package [56] by using F3X4 condon frequency. The input super CDS sequence
494 was from the linked coding sequences (CDS) as used in construct phylogenetic tree of
495 *Flaveria* genus [21], which contains 2462 genes. Gene expression divergence was
496 calculated as Euclidean distance applying R package based on gene expression values
497 (FPKM) of 1,2218 genes. Encoded morphology values of 30 morphology traits were
498 from [41]. The morphology divergence was calculated as Euclidean distance of
499 morphology values. Expression and morphology values were normalized using
500 quantile normalization applying preprocessCore package in R. Linear regression of
501 pair-wise correlation was inferred apply lm function in R package.

502 **Relative difference of each ancestral node in the phylogenetic tree**

503 The morphological characteristics, gene expression abundance, and protein
504 sequences at the whole transcriptomic scale were predicted using FASTML [57]. The
505 protein alignment was from [21]. Gene expression abundance and morphological
506 characteristics of all ancestral nodes were predicted by applying ape package of R
507 which uses a maximal likelihood method. For all C₄ related gene expression, protein

508 sequences and physiological data, their values of the ancestral nodes were assigned to
509 those of the most recent species derived from the node.

510 Relative difference of protein sequence at each ancestral node was inferred by
511 comparing the sequence at this node (N) with the nearest preceding node of N (N[pre]),
512 e.g., the number of different amino acid between node2 (N2) with N1 is the number of
513 changed amino acid at N2. The number of different amino acid changes divided by the
514 aligned length of the protein was calculated as relative protein difference for each gene.
515 Relative difference of gene expression and morphology were calculated as $(N$
516 $-N[pre])/N[pre]$. In most cases, the nearest preceding node of N[i] is N[i-1], there are
517 two exceptions: the ancestral node of N11 is N5, and N10 is N8 (Fig. 7D). One-way
518 ANOVA analysis followed by Tukey's Post Hoc test was used to calculate the
519 significance of relative difference between any two ancestral nodes. *P* values were
520 adjusted by *Benjamin-Hochberg* (BH) correction.

521 **List of Abbreviations**

522 A: CO₂ assimilation rate; AAP6: protein and amino acid permease 6; AlaAT:
523 Alanine aminotransferase; AspAT5: aspartate aminotransferase 5; BSCs: bundle sheath
524 cells; CET: cyclic electron transport; Cytb561: cytochrome b561; DE: differentially
525 expressed; FPKM: fragments per kilobase of transcript per million mapped reads; GDC:
526 glycine decarboxylase complex; GLYK: glycerate kinase; GOGAT: glutamine
527 synthetase and glutamine oxoglutarate aminotransferase; GR: glutathione reductase;
528 GSL1: glutamine synthetase-like 1; Lhca5: PSI light harvesting complex gene 5;

529 MATE: multidrug and toxic compound extrusion; MCs: mesophyll cells; NADP-ME:
530 NADP-dependent malic enzyme; NCBI: National Center for Biotechnology
531 Information; Ndh: NADH dehydrogenase-like; NHD1: sodium: hydrogen (Na⁺/H⁺)
532 antiporter 1; PEPC: phosphoenolpyruvate carboxylase; PIFI : post-illumina
533 chlorophyll fluorescence increase; PNUE: instantaneous photosynthetic nitrogen use
534 efficiency; PPCKA: PEPC protein kinase A; PPDK-RP: PPDK regulatory protein;
535 PPDK: pyruvate, orthophosphate dikinase; PSI: photosystem I; PWUE: instantaneous
536 photosynthetic water use efficiency; RAC: Rubisco activase; Rubisco:
537 ribulose-1,5-bisphosphate carboxylase/oxygenase; SHM: hydroxymethyltransferase;
538 SorDH: sorbitol dehydrogenase; SRA: Sequence Read Achieve; Γ : CO₂ compensation
539 point;

540 **Acknowledgement**

541 The authors thank Haiyang Hu for great suggestion, and thank Chinese Academy
542 of Sciences and the Max Planck Society for support. This work was sponsored by
543 Shanghai Sailing Program [17YF421900], National Science Foundation of China
544 [31701139 to Ming-Ju Amy Lyu, 30970213 to Xin-Guang Zhu]; Bill & Melinda Gates
545 Foundation [OPP1014417 to Xin-Guang Zhu]. We thank LetPub (www.letpub.com) for
546 its linguistic assistance during the preparation of this manuscript. All authors declared
547 not conflict of interest.

548 **Competing interests**

549 None of the authors have any competing interests.

550 **Availability of data and materials**

551 RNA-Seq used in this study is downloaded from Sequence Read Achieve (SRA)
552 of the National Center for Biotechnology Information (NCBI), the accession number is
553 showed in Table S1.

554 **Authors' contributions**

555 MAL, UG, YT, HC and SC conducted the analysis and wrote the paper. PW, SK,
556 JMH, RFS, ML, GKS and XGZ designed the study and wrote the paper.
557

Reference

1. Pattin KA, Moore JH: **Genome-wide association studies for the identification of biomarkers in metabolic diseases.** *Expert opinion on medical diagnostics* 2010, **4**(1):39-51.
2. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A *et al*: **Finding the missing heritability of complex diseases.** *Nature* 2009, **461**(7265):747-753.
3. Huang X, Han B: **Natural variations and genome-wide association studies in crop plants.** *Annual review of plant biology* 2014, **65**:531-551.
4. Zhu X-G, Shan L, Wang Y, Quick WP: **C4 Rice - an ideal arena for systems biology research.** *Journal of Integrative Plant Biology* 2010, **52**(8):762-770.
5. Sage RF, Christin PA, Edwards EJ: **The C-4 plant lineages of planet Earth.** *J Exp Bot* 2011, **62**(9):3155-3169.
6. Hatch MD: **C4 photosynthesis: a unique blend of modified biochemistry, anatomy and ultrastructure.** *Biochimica et Biophysica Acta* 1987, **895**:81-106.
7. Sage RF: **The evolution of C4 photosynthesis.** *New Phytologist* 2003, **161**(2):341-347.
8. 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.** *The Biochemical journal* 1968, **106**(1):141-146.
9. 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.** *The Biochemical journal* 1969, **114**(1):127-134.
10. Dengler N, Nelson T: **Leaf structure and development in C4 plants.** In.: Sage, R, F., Monson, R, K ed (s). *C4 plant biology.*. Academic Press: San Diego, etc; 1999.
11. Hatch MD, Osmond CB: **Compartmentation and transport in C4 photosynthesis.** *Encyclopedia of Plant Physiology* 1976, **3**:144-184.
12. Slack CR, Hatch MD, Goodchild DJ: **Distribution of enzymes in mesophyll**

- 591 **and parenchyma-sheath chloroplasts of maize leaves in relation to the**
- 592 **C4-dicarboxylic acid pathway of photosynthesis. *The Biochemical***
- 593 *journal* 1969, **114**(3):489-498.
- 594 13. Sage RF, Sage TL, Kocacinar F: **Photorespiration and the evolution of C4**
- 595 **photosynthesis. *Annual review of plant biology* 2012, **63**:19-47.**
- 596 14. McKown AD, Dengler NG: **Key innovations in the evolution of Kranz**
- 597 **anatomy and C4 vein pattern in Flaveria (Asteraceae). *American***
- 598 *journal of botany* 2007, **94**(3):382-399.
- 599 15. Brown NJ, Parsley K, Hibberd JM: **The future of C4 research--maize,**
- 600 **Flaveria or Cleome? *Trends in plant science* 2005, **10**(5):215-221.**
- 601 16. Gowik U, Brautigam A, Weber KL, Weber AP, Westhoff P: **Evolution of C4**
- 602 **photosynthesis in the genus Flaveria: how many and which genes**
- 603 **does it take to make C4? *Plant Cell* 2011, **23**(6):2087-2105.**
- 604 17. Mallman J, Heckmann D, Brautigam A, Lercher MJ, Webb APM, Westhoff P,
- 605 Gowik U: **The role of photorespiration during the evolution of C4**
- 606 **photosynthesis in the genus Flaveria. *eLife* 2014, **3**:e02478.**
- 607 18. Engelmann S, Blasing OE, Gowik U, Svensson P, Westhoff P: **Molecular**
- 608 **evolution of C4 phosphoenolpyruvate carboxylase in the genus**
- 609 **Flaveria--a gradual increase from C3 to C4 characteristics. *Planta***
- 610 2003, **217**(5):717-725.
- 611 19. Westhoff P, Gowik U: **Evolution of c4 phosphoenolpyruvate**
- 612 **carboxylase. Genes and proteins: a case study with the genus Flaveria.**
- 613 *Annals of botany* 2004, **93**(1):13-23.
- 614 20. Engelmann S, Wiludda C, Burscheidt J, Gowik U, Schlue U, Koczor M,
- 615 Streubel M, Cossu R, Bauwe H, Westhoff P: **The gene for the P-subunit of**
- 616 **glycine decarboxylase from the C4 species Flaveria trinervia: analysis**
- 617 **of transcriptional control in transgenic Flaveria bidentis (C4) and**
- 618 **Arabidopsis (C3). *Plant physiology* 2008, **146**(4):1773-1785.**
- 619 21. Lyu MJ, Gowik U, Kelly S, Covshoff S, Mallmann J, Westhoff P, Hibberd JM,
- 620 Stata M, Sage RF, Lu H *et al*: **RNA-Seq based phylogeny recapitulates**
- 621 **previous phylogeny of the genus Flaveria (Asteraceae) with some**
- 622 **modifications. *BMC evolutionary biology* 2015, **15**(1):116.**
- 623 22. Edwards GE, Ku MS: **Biochemistry of C3-C4 intermediates. In: Hatch**

- 624 **MD, Boardman NK, editors. The biochemistry of plants New York:**
625 *Academic Press* 1987:275-325.
- 626 23. Ku MS, Wu J, Dai Z, Scott RA, Chu C, Edwards GE: **Photosynthetic and**
627 **photorespiratory characteristics of flaveria species.** *Plant physiology*
628 1991, **96**(2):518-528.
- 629 24. Matsuoka M, Tada Y, Fujimura T, Kano-Murakami Y: **Tissue-specific**
630 **light-regulated expression directed by the promoter of a C4 gene,**
631 **maize pyruvate,orthophosphate dikinase, in a C3 plant, rice.**
632 *Proceedings of the National Academy of Sciences of the United States of*
633 *America* 1993, **90**(20):9586-9590.
- 634 25. Schaffner AR, Sheen J: **Maize C4 photosynthesis involves differential**
635 **regulation of phosphoenolpyruvate carboxylase genes.** *The Plant*
636 *journal : for cell and molecular biology* 1992, **2**(2):221-232.
- 637 26. Li Y, Xu J, Haq NU, Zhang H, Zhu XG: **Was low CO2 a driving force of C4**
638 **evolution: Arabidopsis responses to long-term low CO2 stress.** *J Exp*
639 *Bot* 2014, **65**(13):3657-3667.
- 640 27. Aubry S, Kelly S, Kumpers BM, Smith-Unna RD, Hibberd JM: **Deep**
641 **evolutionary comparison of gene expression identifies parallel**
642 **recruitment of trans-factors in two independent origins of C4**
643 **photosynthesis.** *PLoS genetics* 2014, **10**(6):e1004365.
- 644 28. John CR, Smith-Unna RD, Woodfield H, Covshoff S, Hibberd JM:
645 **Evolutionary convergence of cell-specific gene expression in**
646 **independent lineages of C4 grasses.** *Plant physiology* 2014,
647 **165**(1):62-75.
- 648 29. Chang YM, Liu WY, Shih AC, Shen MN, Lu CH, Lu MY, Yang HW, Wang TY,
649 Chen SC, Chen SM *et al*: **Characterizing regulatory and functional**
650 **differentiation between maize mesophyll and bundle sheath cells by**
651 **transcriptomic analysis.** *Plant physiology* 2012, **160**(1):165-177.
- 652 30. Tausta SL, Li P, Si Y, Gandotra N, Liu P, Sun Q, Brutnell TP, Nelson T:
653 **Developmental dynamics of Kranz cell transcriptional specificity in**
654 **maize leaf reveals early onset of C4-related processes.** *Journal of*
655 *experimental botany* 2014, **65**(13):3543-3555.
- 656 31. Ferro M, Salvi D, Brugiére S, Miras S, Kowalski S, Louwagie M, Garin J,

- 657 Joyard J, Rolland N: **Proteomics of the chlroplast envelope membranes**
658 **from Arabidopsis thaliana.** *Molecular & Cellular Proteomics* 2003,
659 2(5):325-345.
- 660 32. Das S, Pathak R, Choudhury D, Raghuram N: **Genomewide computational**
661 **analysis of nitrate response elements in rice and Arabidopsis.** *Mol*
662 *Genet Genomics* 2007, 278(5):519-525.
- 663 33. Rentsch D, Hirner B, Schmelzer E, Frommer WB: **Salt stress-induced**
664 **proline transporters and salt stress-repressed broad specificity**
665 **amino acid permeases identified by suppression of a yeast amino**
666 **acid permease-targeting mutant.** *The Plant Cell Online* 1996,
667 8(8):1437-1446.
- 668 34. Okumoto S, Schmidt R, Tegeder M, Fischer WN, Rentsch D, Frommer WB,
669 Koch W: **High Affinity Amino Acid Transporters Specifically**
670 **Expressed in Xylem Parenchyma and Developing Seeds of**
671 **Arabidopsis.** *Journal of Biological Chemistry* 2002, 277(47):45338-45346.
- 672 35. Hunt E, Gattolin S, Newbury HJ, Bale JS, Tseng H-M, Barrett DA, Pritchard J:
673 **A mutation in amino acid permease AAP6 reduces the amino acid**
674 **content of the Arabidopsis sieve elements but leaves aphid herbivores**
675 **unaffected.** *J Exp Bot* 2010, 61(1):55-64.
- 676 36. Allen JF, Pfannschmidt T: **Balancing the two photosystems:**
677 **photosynthetic electron transfer governs transcription of reaction**
678 **centre genes in chloroplasts.** *Philos Trans R Soc Lond Ser B-Biol Sci* 2000,
679 355(1402):1351-1357.
- 680 37. Foyer C, Noctor G: **Redox regulation in photosynthetic organisms:**
681 **signaling, acclimation, and practical implications.** *Antioxid Redox*
682 *Signal* 2009, 11(4):861-905.
- 683 38. Vogan PJ, Sage RF: **Water-use efficiency and nitrogen-use efficiency of**
684 **C(3) -C(4) intermediate species of Flaveria Juss. (Asteraceae).** *Plant,*
685 *cell & environment* 2011, 34(9):1415-1430.
- 686 39. Rumpho ME, Ku MS, Cheng SH, Edwards GE: **Photosynthetic**
687 **Characteristics of C(3)-C(4) Intermediate Flaveria Species : III.**
688 **Reduction of Photorespiration by a Limited C(4) Pathway of**
689 **Photosynthesis in Flaveria ramosissima.** *Plant physiology* 1984,

- 690 75(4):993-996.
- 691 40. Nakamura N, Iwano M, Havaux M, Yokota A, Munekage YN: **Promotion of**
692 **cyclic electron transport around photosystem I during the evolution**
693 **of NADP-malic enzyme-type C4 photosynthesis in the genus Flaveria.**
694 *The New phytologist* 2013, **199**(3):832-842.
- 695 41. McKown AD, Moncalvo J-M, Dengler NG: **Phylogeny of Flaveria**
696 **(Asteraceae) and inference of C4 photosynthesis evolution.** *American*
697 *Journal of Botany* 2005, **92**(11):1911-1928.
- 698 42. Sage RF, Zhu X-G: **Exploiting the engine of C4 photosynthesis.** *Journal of*
699 *experimental botany* 2011, **62**(9):2989-3000.
- 700 43. Wang Y, Long SP, Zhu XG: **Elements required for an efficient**
701 **NADP-malic enzyme type C4 photosynthesis.** *Plant physiology* 2014,
702 **164**(4):2231-2246.
- 703 44. Wedding RT, Black MK, Meyer CR: **Inhibition of phosphoenolpyruvate**
704 **carboxylase by malate.** *Plant physiology* 1990, **92**(2):456-461.
- 705 45. Wang S, Tholen D, Zhu XG: **C4 photosynthesis in C3 rice: a theoretical**
706 **analysis of biochemical and anatomical factors.** *Plant, cell &*
707 *environment* 2017, **40**(1):80-94.
- 708 46. Hunt BG, Ometto L, Keller L, Goodisman MAD: **Evolution at Two Levels in**
709 **Fire Ants: The Relationship between Patterns of Gene Expression and**
710 **Protein Sequence Evolution.** *Mol Biol Evol* 2013, **30**(2):263-271.
- 711 47. Warnefors M, Kaessmann H: **Evolution of the Correlation between**
712 **Expression Divergence and Protein Divergence in Mammals.** *Genome*
713 *Biol Evol* 2013, **5**(7):1324-1335.
- 714 48. Morgan CL, Turner SR, Rawsthorne S: **Coordination of the Cell-Specific**
715 **Distribution of the 4 Subunits of Glycine Decarboxylase and of Serine**
716 **Hydroxymethyltransferase in Leaves of C3-C4 Intermediate Species**
717 **from Different Genera.** *Planta* 1993, **190**(4):468-473.
- 718 49. Sage TL, Busch FA, Johnson DC, Friesen PC, Stinson CR, Stata M, Sultmanis
719 S, Rahman BA, Rawsthorne S, Sage RF: **Initial events during the**
720 **evolution of C4 photosynthesis in C3 species of Flaveria.** *Plant*
721 *physiology* 2013, **163**(3):1266-1276.
- 722 50. Moore Bd, Ku M, S. B, Edwards G, E.: **C4 photosynthesis and**

723 **light-dependent accumulation of inorganic carbon in leaves of C3-C4**
724 **and C4 Flaveria species.** *Australian Journal of Plant Physiology* 1987,
725 **14:**658-668.

726 51. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis
727 X, Fan L, Raychowdhury R, Zeng Q *et al*: **Full-length transcriptome**
728 **assembly from RNA-Seq data without a reference genome.** *Nature*
729 *biotechnology* 2011, **29**(7):644-652.

730 52. Huang X, Madan A: **CAP3: A DNA sequence assembly program.** *Genome*
731 *research* 1999, **9**(9):868-877.

732 53. Li B, Dewey CN: **RSEM: accurate transcript quantification from**
733 **RNA-Seq data with or without a reference genome.** *BMC bioinformatics*
734 2011, **12**:323.

735 54. Brawand D, Soumillon M, Necsulea A, Julien P, Csardi G, Harrigan P, Weier
736 M, Liechti A, Aximu-Petri A, Kircher M *et al*: **The evolution of gene**
737 **expression levels in mammalian organs.** *Nature* 2011,
738 **478**(7369):343-348.

739 55. Klambauer G, Unterthiner T, Hochreiter S: **DEXUS: identifying differential**
740 **expression in RNA-Seq studies with unknown conditions.** *Nucleic acids*
741 *research* 2013, **41**(21):e198.

742 56. Yang Z: **PAML: a program package for phylogenetic analysis by**
743 **maximum likelihood.** *Computer applications in the biosciences : CABIOS*
744 1997, **13**(5):555-556.

745 57. Ashkenazy H, Penn O, Doron-Faigenboim A, Cohen O, Cannarozzi G, Zomer
746 O, Pupko T: **FastML: a web server for probabilistic reconstruction of**
747 **ancestral sequences.** *Nucleic acids research* 2012, **40**(Web Server
748 issue):W580-584.

749
750

751 **Supplementary Information:**

752 **Additional file 1:** Includes supplemental methods, figures and tables

753 **Additional file 2:** The alignments of proteins

754 **Additional file 3:** 205 genes that showed differential expression and at least one amino

755 acid change between C₃ and C₄ species. The table includes the gene identifier,

756 expression fold-change and number of amino acid changes between C₄ and C₃ species.

757 **Additional file4:** FPKM of all genes

758

759 Figure legends

760 **Figure 1. Modifications in phosphoenolpyruvate carboxylase and pyruvate** 761 **orthophosphate dikinase predicted protein sequences and transcript abundances** 762 **mapped to the *Flaveria* phylogeny.**

763 The predicted amino acid changes in phosphoenolpyruvate carboxylase (PEPC) and
764 pyruvate orthophosphate dikinase (PPDK) between C₄ and C₃ *Flaveria* species and the
765 transcript abundance (FPKM) of the genes encoding the proteins are shown. Only the
766 amino acid residues predicted to be different between C₃ and C₄ species are
767 superimposed on the schema of *Flaveria* phylogenetic tree modified from Lyu *et al.*,
768 2015. The colors of amino acid residues have no meaning and are only for visualization
769 purposes. Numbers below the amino acids indicate the location sites in the multiple
770 sequence alignments. FPKM values are shown to the right of the amino acid changes as
771 red bars. A: PEPC. B: PPDK. Protein sequences from UniprotKP are: *F. trinervia* PEPC,
772 P30694; *F. bidentis* PPDK, Q39735; *F. brownii* PPDK, Q39734; and *F. trinervia* PPDK,
773 P22221. Sequence alignments are available in Additional file 2.

775 **Figure 2. Modifications in NADP-malic enzyme, pyruvate orthophosphate** 776 **dikinase regulatory protein and phosphoenolpyruvate protein kinase A predicted** 777 **protein sequences and transcript abundances mapped to the *Flaveria* phylogeny.**

778 The predicted amino acid changes in NADP-malic enzyme (NADP-ME), pyruvate
779 orthophosphate dikinase regulatory protein (PPDK-RP) and phosphoenolpyruvate
780 protein kinase A (PPCKA) between C₄ and C₃ species and the transcript abundance
781 (FPKM) of the genes encoding the proteins are shown. Only the amino acid residues
782 predicted to be different between C₃ and C₄ species are superimposed on the schema of
783 *Flaveria* phylogenetic tree modified from Lyu *et al.*, 2015. The colors of amino acid
784 residues have no meaning and are only for visualization purposes. Numbers below the
785 amino acids indicate the location site in the multiple sequence alignments. FPKM
786 values are represented to the right of the amino acid changes as red bars. A: NADP-ME.
787 B: PPDK-RP. C: PPCKA. The sequence alignments are available in Additional file 2.

789 **Figure 3. Modifications in photorespiratory protein predicted amino acid**

790 **sequences and cognate transcript abundances mapped to the *Flaveria* phylogeny.**

791 The predicted amino acid changes in photorespiratory proteins between C₄ and C₃
792 *Flaveria* species and the transcript abundance (FPKM) of genes encoding the proteins
793 are shown. Only the amino acid residues that are predicted to be different between C₃
794 and C₄ species are superimposed on the schema of *Flaveria* phylogenetic tree modified
795 from Lyu *et al.*, 2015. The marked colors of amino acid residues have no meaning and
796 are only for visualization purposes. Numbers below the amino acids indicate the
797 location site in the multiple sequence alignments. FPKM values are represented to the
798 right of the amino acid changes as red bars. A: glycine decarboxylase complex H
799 subunit (GDC-H); B: serine hydroxymethyltransferase (SHM); C: glycerate kinase
800 (GLYK); D: glutamine synthetase and glutamine oxoglutarate aminotransferase
801 (GOGAT); E, Glutamine synthetase like 1 (GSL1).

802

803 **Figure 4. Modifications in the predicted amino acid sequences of proteins**
804 **involved in cyclic electron transport and transcript abundances of the cognate**
805 **transcripts mapped to the *Flaveria* phylogeny.**

806 Changes in predicted amino acid sequence in proteins involved in cyclic electron
807 transport and abundances (FPKM) of their cognate transcripts in C₄ and C₃ *Flaveria*
808 species are shown. Only the amino acid residues predicted to be different between C₃
809 and C₄ species are superimposed on the schema of *Flaveria* phylogenetic tree modified
810 from Lyu *et al.*, 2015. The marked colors of amino acid residues have no meaning and
811 are only for visualization purposes. Numbers below the amino acids indicate the
812 location site in the multiple sequence alignments. FPKM values are represented to the
813 right of the amino acid changes as red bars. A: protein gradient regulation 5 like protein
814 (PGR5-like); B: NADH dehydrogenase-like (Ndh) L2 subunit (Ndh L2); C: NdhV; D:
815 Ndh16; E: NdhU; F: NdhM; G: Ndh48; H: NdhB4; I: chlororespiration reduction 1.
816 The sequence alignments are available in Additional file 2.

817

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

820 Overall, C₄-related physiological (green and blue bars) and anatomical traits (orange
821 and red bars) showed a step-wise change along the *Flaveria* phylogenetic tree; however,
822 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 : CO₂ compensation point; A: CO₂ assimilation rate; PWUE: instantaneous photosynthetic water use efficiency; PNUE: instantaneous photosynthetic nitrogen use efficiency; response slope: slope of the response of net CO₂ 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 Mckown 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 according to the evolutionary time. *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.

Figure 7. Evolutionary pattern of C₄ related genes comparing with total expressed genes and DE genes between C₃ and C₄ species in gene expression

There dataset are (A) the total expressed gene (12215 genes), (B) differentially expressed genes between C₃ and C₄ species (2306 genes) and (C) C₄ related genes that showed difference between C₃ and C₄ species in both protein sequence and gene expression (205 genes). All the three datasets showed that the biggest change of gene expression occurred at N7 (left panels). Principle Component Analysis (PCA) showed that species derived from N7 (species in the black frames in right panels) are distinguished with other species, and the first component of the C₄ related genes

856 account for 38% of the total variance, more than that of total expressed genes and DE
857 genes (right panels). The bar colors grey/blue/orange represent: species from
858 basal/clade A/clade B of phylogenetic tree.
859
860

861 Tables

862 **Table 1. Proteins showing differences in amino acid sequence between C₃ and C₄**
 863 ***Flaveria* species and the relative changes in their cognate transcripts**

Ortholog in <i>A. thaliana</i>	Genes encoding proteins involved in	Mean FPKM (C ₄)/mean FPKM (C ₃)	Length in Fcro (Frob) ^a	Protein length in <i>A. thaliana</i> (aa)	aa changes						Stage of key change(s) in sequence	Stage of key change(s) in FPKM ^b
					total aa change(s)	before N 5	at N5	at N6	at N7	after N7		
Gene in C4 pathway												
AT3G14940	PEPC	85.58	966	968	41			1	≥34		N7	N7
AT4G15530	PPDK	123.6	958	963	31			2 + 6-aa REP	≥15		N7	N7
AT1G79750	NADP-ME	26.64	647	646	27		1	8	18		N7	N3 and N6
AT4G21210	PPDK-RP	7.57	402	403	13	1		4	7		N7	N7
AT3G04530	PEPC-k	88.78	281	278	12	3		2	7		N7	N7
AT1G72330	AlaAT	9.63	544	553	9			2	7		N7	N3 an N6
AT4G31990	AspAT5	36.67	459	453	3			1	1	1	N7	N3 and N7
AT2G26900	BASS2	39.12	415	409	14			2	12		N7	N7
AT3G19490	NHD1	51.19	576	576	15			2	13		N7	N7
Gene in photorespiration pathway												
AT1G32470	GDC-H	0.23	162	166	6+2-aa INS + 1-aa INS				5 + 2-aa INS + 1-aa INS	1	N7	N7
AT4G37930	SHM	0.16	517	517	8	3			5	1	N7	N7
AT1G80380	GLYK	0.49	443	456	8			2	6	1	N7	N7
AT5G04140	GOGAT	0.57	1616	1648	18	4		≥1	≥5	2	N7	N7
AT5G35630	GSL1	0.08	430	430	8	3		1	2	1	N7	N7
Gene related to electron transport chain												
AT4G22890	PGR5-like	7.1	328	324	10+17-aa INS	1		2+17-aa INS	7		N6	N7
AT1G14150	NdhL2/PnsL2	3.71	190	190	4	2		1	1		before N5	N7
AT2G04039	NdhV	9.17	227	199	8		1	6	1		N6	N3
AT5G43750	Ndh18/PnsB5	6.8	224	212	3			2	1		N6	N8
AT5G21430	NdhU/CRRL	8.55	215	218	4			4			N6	N7
AT4G37925	NdhM	7.01	209	217	3			1	2		N7	N8
AT1G15980	Ndh48/PnsB1	8.6	465	461	7	1		4	2		N6	N7
AT1G18730	NdhB4/PnsB4	8.8	182	174	5		1	2	2		N6 and N7	N7
AT5G52100	CRR1	4.05	302	298	5			1	1	3	after N7	N3
AT1G45474	Lhca5	13.5	268	256	5		3	1	2		N5	N3
AT3G15840	PIF1	6.78	283	268	10		5	1	4		N5 an N7	N4
AT3G07570	Cytochrome b561	5.99	373	369	6	1	1		3	1	N7	N3 and N4
Photosynthesis												
AT2G39730	RCA	0.47	475	474	11			9	2*		N6	N7
AT5G12470	response to nitrate level	10.03	372 *	386	33 +3-aa INS+2-aa INS			11+3-aa INS	22+2-aa INS		N7	N7
Transport												
AT5G65380	MATE	2.35	496	486	4			3	1		N6	N9
AT5G49630	AAP6	3.52	472	481	4	1			3		N7	N8
Oxidation-reduction												
AT3G54660	GR	2.2	571	565	6	1		2	2	1	N6 and N7	N3
AT5G51970	SorDH	0.31	362	364	3				3		N7	N7

864

865 ^a: protein length in Frob or in Fero if the protein in Frob is not completely assembled. ^b:

866 The stages of key change in FPKM of each gene were determined based on the relative
867 change of each node as showed in Fig. S12. *: incomplete assembled sequence.

868 Abbreviations: aa: amino acid, Frob: *F. robusta*, Fcro: *F. cronquistii*, INS: insertion,
869 DEL: deletion, REP: replacement.

870

871

Figure 2. Modifications in NADP-malic enzyme, pyruvate orthophosphate dikinase regulatory protein and phosphoenolpyruvate protein kinase A predicted protein sequences and transcript abundances mapped to the *Flaveria* phylogeny.

The predicted amino acid changes in NADP-malic enzyme (NADP-ME), pyruvate orthophosphate dikinase regulatory protein (PPDK-RP) and phosphoenolpyruvate protein kinase A (PPCKA) between C₄ and C₃ species and the transcript abundance (FPKM) of the genes encoding the proteins are shown. Only the amino acid residues predicted to be different between C₃ and C₄ 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 site in the multiple sequence alignments. FPKM values are represented to the right of the amino acid changes as red bars. A: NADP-ME. B: PPK-RP. C: PPKA. The sequence alignments are available in Additional file 2.

Figure 4. 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 and abundances (FPKM) of their cognate transcripts in C₄ and C₃ *Flaveria* species are shown. Only the amino acid residues predicted to be different between C₃ and C₄ 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 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. The sequence alignments are available in Additional file 2.

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 Mckown 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 according to the evolutionary time. 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.

Figure 7. Evolutionary pattern of C₄ related genes comparing with total expressed genes and DE genes between C₃ and C₄ species in gene expression

There dataset are (A) the total expressed gene (8004 genes), (B) differentially expressed genes between C₃ and C₄ species (2306 genes) and (C) C₄ related genes that showed difference between C₃ and C₄ species in both protein sequence and gene expression (205 genes). All the three datasets showed that the biggest change of gene expression occurred at N7 (left panels). Principle Component Analysis (PCA) showed that species derived from N7 (species in the black frames in right panels) are distinguished with other species, and the first component of the C₄ related genes account for 38% of the total variance, more than that of total expressed genes and DE genes (right panels). The bar colors grey/blue/orange represent: species from basal/clade A/clade B of phylogenetic tree.