Integrated genomic analysis revealed a dominant role of transcriptional regulation during the evolution of C_4 photosynthesis in *Flaveria* species

Ming-Ju Amy Lyu[1,9], Huilong Du[2,3,9], Hongyan Yao[4,9], Zhiguo Zhang[5,9], Genyun Chen[1], Yuhui Huang[1,6], Xiaoxiang Ni[1,6], Faming Chen[1,6], Yong-Yao Zhao[1,6], Qiming Tang[1,6], Fenfen Miao[1,6], Yanjie Wang[1,6], Yuhui Zhao[2], Hongwei Lu[2], Lu Fang[2], Qiang Gao[2], Yiying Qi[7], Qing Zhang[7], Jisen Zhang[7], Tao Yang[8], Xuean Cui[5], Chengzhi Liang[2,3$], Tiegang Lu[5$], Xin-Guang Zhu[1,10$]

[1]State Key Laboratory of Plant Molecular Genetics, Center of Excellence for Molecular Plant Sciences, Chinese Academy of Sciences, Shanghai, China, 200032
[2]State Key Laboratory of Plant Genomics, Institute of Genetics and Developmental Biology, Innovation Academy for Seed Design, Chinese Academy of Sciences, Beijing, China;
[3]University of Chinese Academy of Sciences, Beijing, China; School of Life Sciences, Institute of Life Sciences and Green Development, Hebei University, Baoding, China.
[4]State Key Laboratory of Genetic Engineering, School of Life Sciences, Fudan University, Shanghai 200438, China
[5]Biotechnology Research Institute/National Key Facility for Gene Resources and Gene Improvement, Chinese Academy of Agricultural Sciences, Beijing, 100081, China
6. University of Chinese Academy of Sciences, Beijing 100049, China
7. Center for Genomics and Biotechnology, Fujian Provincial Key Laboratory of Haixia Applied Plant Systems Biology, Key Laboratory of Sugarcane Biology and Genetic Breeding, National Engineering Research Center for Sugarcane, College of Life Sciences, Fujian Agriculture and Forestry University, Fuzhou, China
8. China National GeneBank, Shenzhen, 518120, China.
9. These authors contributed equally
10. Lead Contact
* Correspondence: zhuxg@cemps.ac.cn (X.G.Z), lutiegang@caas.cn (L.T.), cliang@genetics.ac.cn (C.L.)

Supplemental Notes

Table of content

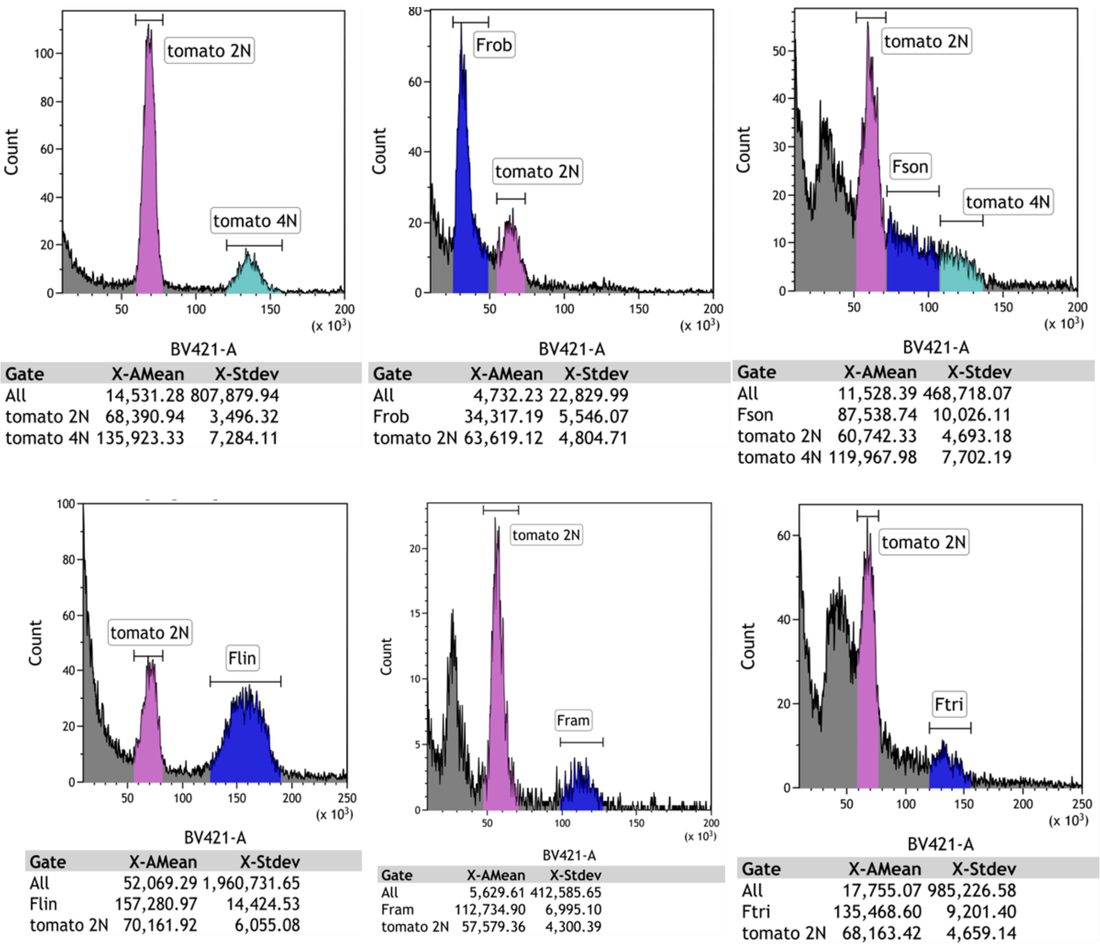
1. Estimation of genome sizes of five *Flaveria* species using flow cytometry

Method

Genome size was estimated using flow-cytometry (FCM). For the FCM, *Solanum lycopersicum* (tomato, genome size 828 MB) was used as reference. Young leaves from multiple species were chopped with a blade soaked in nuclear isolation and staining buffer DAPI (NPE Analyzer, NPE 731085, USA). The suspension was filtered through a 40-lm nylon mesh (Sysmex Cell Trics). Samples were analyzed with a MD FACS Melody flow cytometer and data was analyzed using Kaluza software.

Results

The genome size of the five *Flaveria* species was estimated to be 0.45G for Frob, 1.2G for Fson, 1.86G for Flin, 1.62G for Fram and 1.65G for and Ftri (Fig. S 1)



Gate	X-AMean	X-Stdev
All	14,531.28	807,879.94
tomato 2N	68,390.94	3,496.32
tomato 4N	135,923.33	7,284.11

Gate	X-AMean	X-Stdev
All	4,732.23	22,829.99
Frob	34,317.19	5,546.07
tomato 2N	63,619.12	4,804.71

Gate	X-AMean	X-Stdev
All	11,528.39	468,718.07
Fson	87,538.74	10,026.11
tomato 2N	60,742.33	4,693.18
tomato 4N	119,967.98	7,702.19

Gate	X-AMean	X-Stdev
All	52,069.29	1,960,731.65
Flin	157,280.97	14,424.53
tomato 2N	70,161.92	6,055.08

Gate	X-AMean	X-Stdev
All	5,629.61	412,585.65
Fram	112,734.90	6,995.10
tomato 2N	57,579.36	4,300.39

Gate	X-AMean	X-Stdev
All	17,755.07	985,226.58
Ftri	135,468.60	9,201.40
tomato 2N	68,163.42	4,659.14

Tomato genome size (2n) in bp: 827,747,456

Species	X-mean (*Flaveria* 2n)	X-mean (tomato 2n)	Estimated genome size (bp)
Frob (C_3)	34317.19	63619.12	447,039,887
Fson (C_3-C_4)	87538.74	60742.19	1,194,351,222
Flin (C_3-C_4)	157280.97	70161.92	1,857,791,309
Fram (C_3-C_4)	112734.9	57579.36	1,622,608,563
Ftri (C_4)	135468.6	68163.42	1,647,060,239

Fig. S 1. **Estimation of genome size using flow-cytometry**

Solanum lycopersicum (tomato, genome size: 828 MB) was used as a reference. Samples were mixed for measurement as (a) only tomato, (b) Frob and tomato, (c) Fson and tomato, (d) Flin and tomato, (e) Fram and tomato and (f) Ftri and tomato. Young leaves from multiple species were chopped with a blade in nuclear isolation and the staining buffer DAPI. Estimated genome size of *Flaveria* species according to tomato are shown in (g). (Abbreviations: Frob: *F. robusta*; Fson: *F. sonorensis*; Flin: *F. linearis*; Fram: *F. ramosissima*; Ftri: *F. trinervia*)

2. Investigation of chromosome numbers using Fluorescence in situ hybridization assays

Methods

To investigate the chromosome number of Frob, Flin and Ftri, mitotic metaphase spreads were prepared from meristem root tip cells following a previously published method [1] with minor modifications. Briefly, root tips approximately 1 cm in length were cut and pretreated with nitrous oxide gas for 1-3 hours. The root tips were then fixed in ice-cold 90% acetic acid for 10 minutes and stored in 70% ethanol at - 20 °C. Root segments with actively dividing regions were excised and incubated in an enzyme mixture containing 1% (w/w) pectolyase Y-23 (Yakult Pharmaceutical, Tokyo, Japan) and 2% (w/w) cellulose Onozula R-10 (Yakult Pharmaceutical) for 30-50 minutes at 37 °C. After digestion, the root sections were washed twice in 75% ethanol. The root sections were fine-broken with a needle and treated in vortex machine at 2000 rpm for 20 seconds. The cells were collected by centrifugation and re-suspended in 30 ml 100% acetic acid to prepare a cell suspension. The cell suspension (5~8 μL) was placed onto glass slides in a moist box and dried. The slides were cross-linked in an ultraviolet cross-linking instrument and dyed using DAPI (NPE Analyzer, NPE 731085, USA), before being viewed under a microscope (Leica DM2500).

Results

The chromosome number was 2x18 in all the three analyzed species (Fig.1a), in line with previous reports. Additionally, Hi-C assembly supported the chromosome number of 18 for all five *Flaveria* species (Fig. S 1). The synteny of the 18 chromosomes was conserved across the five *Flaveria* species; from 50% to 75% of protein coding genes being colinear between Frob and the other species (Table S 1).

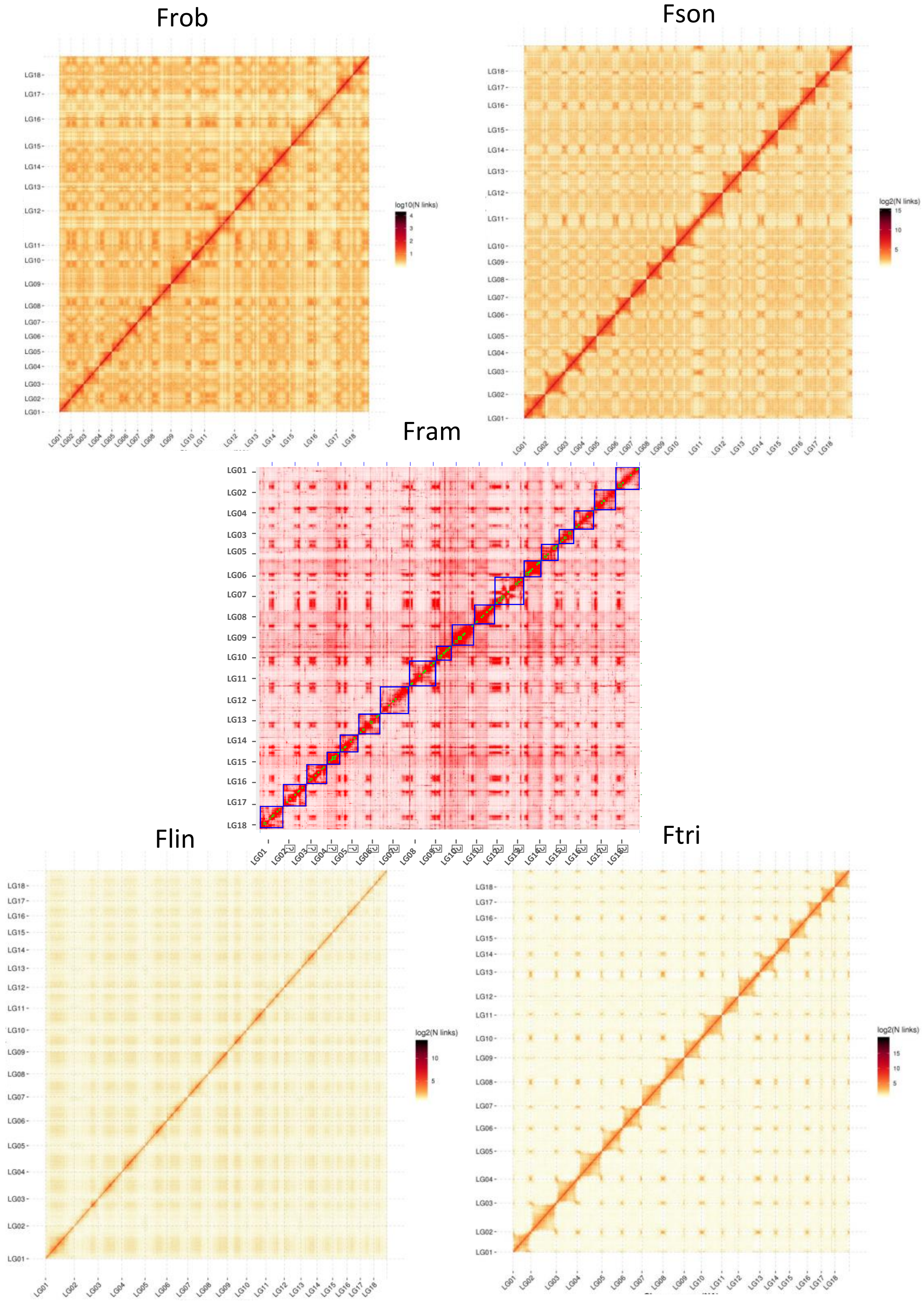


Fig. S 2. **Contact maps of the chromosomes of each *Flaveria* species**

Normalized Hi-C contract matrix and assembled chromosomes number from 1 to 18 are shown in fragments of 100kb. X and Y lab show the chromosomes numbered from1 to 18.

Table S 1. **The proportions of colinear genes between Frob and the other four sequenced *Flaveria* species**

Pair	% Collinearity
Fson *vs*. Frob	72.46
Flin *vs*. Frob	50.15
Fram *vs*. Frob	51.96
Ftri *vs*. Frob	75.37

3. Estimation of genome assembly completeness

Methods:

The completeness of genome assembly was estimated in three methods. First, The completeness of protein repertoire was estimated using Benchmarking Universal Single Copy Orthologues (BUSCO) (v3.0.2) [2] against a viridiplantae reference. Second, the completeness of genome assembly was estimated through RNA-seq mapping. RNA short reads sequencing was performed using the Illumina platform to obtain paired-ended reads with a length of 150 bp. RNA-seq reads from multiple experiments were mapped to the assembled genome sequences applying RSEM [3] in default parameters, where STAR (v2.7.3a) [4] was selected as the mapping tool. Third, the completeness of genome assembly was estimated through DNA-seq short reads mapping. DNA short reads sequencing was performed using the Illumina platform to obtain paired-ended reads with a length of 150 bp. DNA-seq reads were mapped to genome apply bowtie2 (v2.3.4.3) [5] in with “-k 10”.

Results

BUSCO results demonstrated that 99.2% (Frob) to 92.5% (Flin) of BUSCO genes were covered (Table S 2). RNA-seq reads mapping results showed that a total of 3,290 million reads from five independent experiments were used for the five *Flaveria* species, and from 21 to 34 million reads were generated for each sample. For Frob, Fson, Fram and Ftri, between 92.0% and 97.6% of RNA-seq reads were mapped to corresponding genomes. Flin had a relatively lower mapping rates than other species, *i.e.*, around 86.7% to 90%. RNA-seq datasets from Flin from the high light experiment were generated using a different accession of Flin, which had relatively lower mapping rates, *i.e.*, around 80% (Table S 3). DNA-seq reads mapping results indicated that 1,110 million reads in total were used for the five *Flaveria* species, and between 170 and 324 million reads were generated for each species. From 95% to 98% paired-end reads were mapped to corresponding assembled genomes (Table S 4). Therefore, the above results suggested completeness and high quality of the genome assembly.

Table S 2. **Statistics of BUSCO analysis**

Species	**Complete (%)**	**Complete (single) (%)**	**Complete (duplicated) (%)**
Frob (C_3)	99.2	89.6	9.6
Fson (C_3-C_4)	98.1	89.6	8.5
Flin (C_3-C_4)	92.5	86.4	6.1
Fram (C_3-C_4)	95	88.4	6.6
Ftri (C_4)	95.1	91.1	4

Table S 3**. Statistics of RNA-seq mapping**

2-week low CO_2 experiment	**# Total reads**	**Unique mapping ratio (%)**	**Multiple mapping ratio (%)**	**Total mapping ratio (%)**
Flin_2w_L_1$^{@}$	26106017	77.8	8.9	86.7
Flin_2w_L_2	33219838	79.6	9.8	89.4
Flin_2w_L_3	28760467	78.7	8.9	87.6
Flin_2w_N_1	31595547	80.5	9.5	90.0
Flin_2w_N_2	32903828	80.3	9.7	90.0
Flin_2w_N_3	33093266	80.6	9.7	90.3
Fram_2w_L_1	34936336	88.4	3.7	92.1
Fram_2w_L_2	32185750	88.8	3.5	92.3
Fram_2w_L_3	35172051	88.2	3.6	91.8
Fram_2w_N_1	15860328	88.3	3.7	92.0
Fram_2w_N_2	24661281	88.5	3.6	92.2
Fram_2w_N_3	34759285	88.4	3.8	92.2
Frob_2w_L_1	33887554	91.2	5.6	96.7
Frob_2w_L_2	36520872	91.6	5.5	97.1
Frob_2w_L_3	28938440	91.2	5.6	96.8
Frob_2w_N_1	33760952	91.0	5.8	96.8
Frob_2w_N_2	26874262	91.1	5.7	96.8
Frob_2w_N_3	34406245	91.3	5.8	97.1
Fson_2w_L_1	28557889	91.7	4.2	95.9
Fson_2w_L_2	26985392	91.7	4.1	95.8
Fson_2w_L_3	34615832	91.7	4.2	95.9

Fson_2w_N_1	28463981	91.5	4.3	95.8
Fson_2w_N_2	26500228	91.6	4.4	95.9
Fson_2w_N_3	32336172	91.8	4.3	96.1
Ftri_2w_L_1	35188793	92.2	4.3	96.5
Ftri_2w_L_2	29165653	89.0	7.4	96.4
Ftri_2w_L_3	32889783	92.3	4.4	96.6
Ftri_2w_N_1	31460531	92.2	4.0	96.2
Ftri_2w_N_2	32803186	92.5	4.1	96.6
Ftri_2w_N_3	29192442	92.6	3.9	96.5
4-week low CO_2 experiment	**# Total reads**	**Unique mapping ratio (%)**	**Multiple mapping ratio (%)**	**Total mapping ratio (%)**
Flin_4w_L_1	28,713,204	81.7	9.5	91.1
Flin_4w_L_2	23,778,224	81.5	9.4	90.9
Flin_4w_L_3	26,102,203	81.1	9.6	90.7
Flin_4w_N_1	24,809,359	80.9	9.6	90.5
Flin_4w_N_2	24,953,829	80.6	9.9	90.5
Flin_4w_N_3	23,095,900	80.3	9.9	90.1
Fram_4w_L_1	26,122,863	88.9	3.5	92.4
Fram_4w_L_2	19,480,458	89.3	3.6	92.8
Fram_4w_L_3	25,266,411	89.4	3.5	92.9
Fram_4w_N_1	22,702,731	89.3	3.5	92.8
Fram_4w_N_2	25,915,298	89.5	3.5	93.0
Fram_4w_N_3	25,468,944	89.5	3.5	93.0
Frob_4w_L_1	23,432,277	91.5	5.7	97.3
Frob_4w_L_2	24,386,859	90.9	5.7	96.6
Frob_4w_L_3	21,307,320	91.4	5.6	97.0
Frob_4w_N_1	28,024,652	91.2	5.9	97.1
Frob_4w_N_2	26,865,706	91.1	5.9	97.0
Frob_4w_N_3	25,369,429	90.9	5.9	96.8
Fson_4w_L_1	25,504,504	91.8	4.2	96.0
Fson_4w_L_2	16,723,354	91.9	4.3	96.2
Fson_4w_L_3	29,653,035	92.1	4.2	96.3
Fson_4w_N_1	27,538,862	91.3	4.5	95.8
Fson_4w_N_2	26,785,677	91.8	4.3	96.2
Fson_4w_N_3	29,476,651	91.8	4.3	96.2
Ftri_4w_L_1	28,423,320	92.0	4.7	96.8
Ftri_4w_L_2	27,234,665	92.4	4.7	97.0
Ftri_4w_L_3	26,384,930	91.0	4.6	95.6
Ftri_4w_N_1	28,143,172	92.3	4.5	96.8
Ftri_4w_N_2	23,287,687	92.4	4.4	96.8
Ftri_4w_N_3	22,794,859	92.6	4.5	97.1
6-month low CO_2 experiment	**# Total reads**	**Unique mapping ratio (%)**	**Multiple mapping ratio (%)**	**Total mapping ratio (%)**

Fram_6m_L_1	21847410	90.0	3.5	93.5
Fram_6m_L_2	20826620	89.7	3.6	93.2
Fram_6m_L_3	20952186	90.0	3.6	93.6
Fram_6m_L_4	20198206	89.7	3.5	93.3
Fson_6m_L_1	21858279	92.1	4.5	96.7
Fson_6m_L_2	24323857	92.4	4.3	96.8
Fson_6m_L_3	20875634	92.5	4.4	96.9
Fson_6m_L_4	24896827	91.3	4.5	95.8

ABA experiment	**# Total reads**	**Unique mapping ratio (%)**	**Multiple mapping ratio (%)**	**Total mapping ratio (%)**
Fro-ABA-1	22,810,323	92.4	3.7	96.0
Fro-ABA-2	26,320,179	92.4	4.4	96.8
Fro-ABA-3#	23,003,611	92.0	5.6	97.6
Fro-Ctrl-1	26,087,085	92.4	4.0	96.4
Fro-Ctrl-2	21,806,118	91.3	5.1	96.4
Fro-Ctrl-3	34,675,444	92.6	4.5	97.0
Fso-ABA-1	26,905,126	92.0	3.3	95.3
Fso-ABA-2	24,545,444	92.7	4.1	96.7
Fso-ABA-3	25,535,466	91.9	5.9	97.8
Fso-Ctrl-1	26,002,565	92.4	4.3	96.7
Fso-Ctrl-2	28,004,493	92.4	4.5	96.9
Fso-Ctrl-3	25,087,187	91.8	5.1	96.9
Fra-ABA-1	22,157,100	88.4	4.7	93.2
Fra-ABA-3	24,617,078	89.2	3.3	92.5
Fra-ABA-4	22,164,481	88.3	5.6	93.9
Fra-Ctrl-1	28,575,584	88.6	3.4	92.0
Fra-Ctrl-2	27,412,141	88.6	4.3	92.9
Fra-Ctrl-3	26,981,126	89.2	4.5	93.7
Ftr-ABA-1	24,748,259	92.4	3.5	95.9
Ftr-ABA-2	23,080,636	91.7	5.4	97.1
Ftr-ABA-3	25,120,473	91.2	6.0	97.2
Ftr-Ctrl-1	22,970,780	93.0	3.7	96.7
Ftr-Ctrl-2	23,937,914	92.4	3.7	96.1
Ftr-Ctrl-3	26,571,868	92.7	4.4	97.1
Average	26,830,957	89.6	5.1	94.8

High-light experiment &	**# Total reads**	**Unique mapping ratio (%)**	**Multiple mapping ratio (%)**	**Total mapping ratio (%)**
Frob_HL_H_1	32059692	75.9	6.8	82.7
Frob_HL_H_2	29783728	71.9	10.3	82.3
Frob_HL_H_3	31070734	74.9	6.9	81.9
Frob_HL_L_1	32903035	75.6	6.1	81.7
Frob_HL_L_2	30903465	77.4	6.8	84.2
Frob_HL_L_3	31645641	77.9	5.8	83.6

Flin_HL_H_1	33192028	72.0	9.2	81.2
Flin_HL_H_2	24233911	70.6	9.1	79.7
Flin_HL_H_3	22992824	69.2	8.9	78.1
Flin_HL_L_1	30945129	71.7	9.1	80.8
Flin_HL_L_2	20898845	71.2	9.1	80.3
Flin_HL_L_3	36695522	71.6	9.1	80.8
Fram_HL_H_1	33766904	75.4	4.5	80.0
Fram_HL_H_2	34883202	77.1	4.6	81.7
Fram_HL_H_3	34641349	76.3	4.7	81.0
Fram_HL_L_1	36517568	77.3	4.9	82.2
Fram_HL_L_2	40191972	75.6	4.8	80.4
Fram_HL_L_3	33976982	77.3	4.7	82.0
Ftri_HL_H_1	41616357	76.7	4.8	81.5
Ftri_HL_H_2	41633337	77.4	4.5	81.9
Ftri_HL_H_3	42552571	77.0	4.8	81.8
Ftri_HL_L_1	43200463	75.4	4.4	79.8
Ftri_HL_L_2	43456489	77.6	4.8	82.5
Ftri_HL_L_3	37944526	76.1	4.5	80.7

@ Sample with the lowest mapping ratio

# Sample with the highest mapping ratio

& High light experiment was not included for computing average mapping ratio.

Table S 4. **DNA-seq mapping statistics**

Species	**# Total reads**	**# Mapped reads**	**Mapping ratio (%)**	**Properly mapped reads**	**Properly mapping ratio (%)**
Frob (C$_3$)	232,281,182	231,705,950	99.75	227,820,606	98.08
Fson (C$_3$-C$_4$)	189,408,029	188,479,930	99.51	184,350,835	97.33
Flin (C$_3$-C$_4$)	170,714,985	169,980,392	99.57	163,505,538	95.78
Fram (C$_3$-C$_4$)	324,792,136	323,557,926	99.62	309,332,030	95.24
Ftri (C$_4$)	202,443,727	201,073,807	99.32	191,310,924	94.50

4. Comparison of protein-coding genes from Taniguchi's assemblies and our assemblies

Methods

The published cDNA sequences from four *Flaveria* species obtained from Taniguchi's assembly [6]. Open reading frames (ORFs) of cDNA were predicted by applying OrfPredictor [7] using default parameters. Genes with an ORF of no less than 100 amino acids were retained for comparisons in both Taniguchi's assembly and our assembly. For Frob, genes from one assembly (either Taniguchi's assembly or our assembly) with counterparts found in another assembly were predicted using Blastp (v2.2.31+) [8] with an E-value threshold of 0.001. Protein identities were obtained from Blastp outputs.

Results

We compared the predicted protein-coding genes from our assembly with those from Taniguchi's assembly [6]. While more protein coding genes were reported in Taniguchi's assembly, approximately 30.7~35.6% of annotated genes were less than 100 amino acid in protein length [6], while the proportion of those in the five species reported here is ~4% (Fig. S 3). Frob is the only species used in both and Taniguchi's assembly and ours. For Frob, 96.1% of the genes with protein length >=100 amino acids from Taniguchi's assembly were covered in our assembly (Blastp, E-value<0.001). In contrast, 96.3% of the genes with protein length >=100 amino acids in our assembly were also covered by Taniguchi's assembly (Blastp, E-value<0.001, Fig. S 4). Therefore, the annotated protein-coding genes in this study can be considered reliable.

We attempted to incorporate protein annotations from Taniguchi's assemblies into our evolutionary comparison study. However, we encountered challenges as several crucial C_4 enzymes, such as CA1, PEPC1, and NADP-ME4, exhibited no annotations sequences in the C_4 species *F. bidentis* (Fig. S 5). Consequently, we opted to exclude the comparison study involving Taniguchi's data in this analysis.

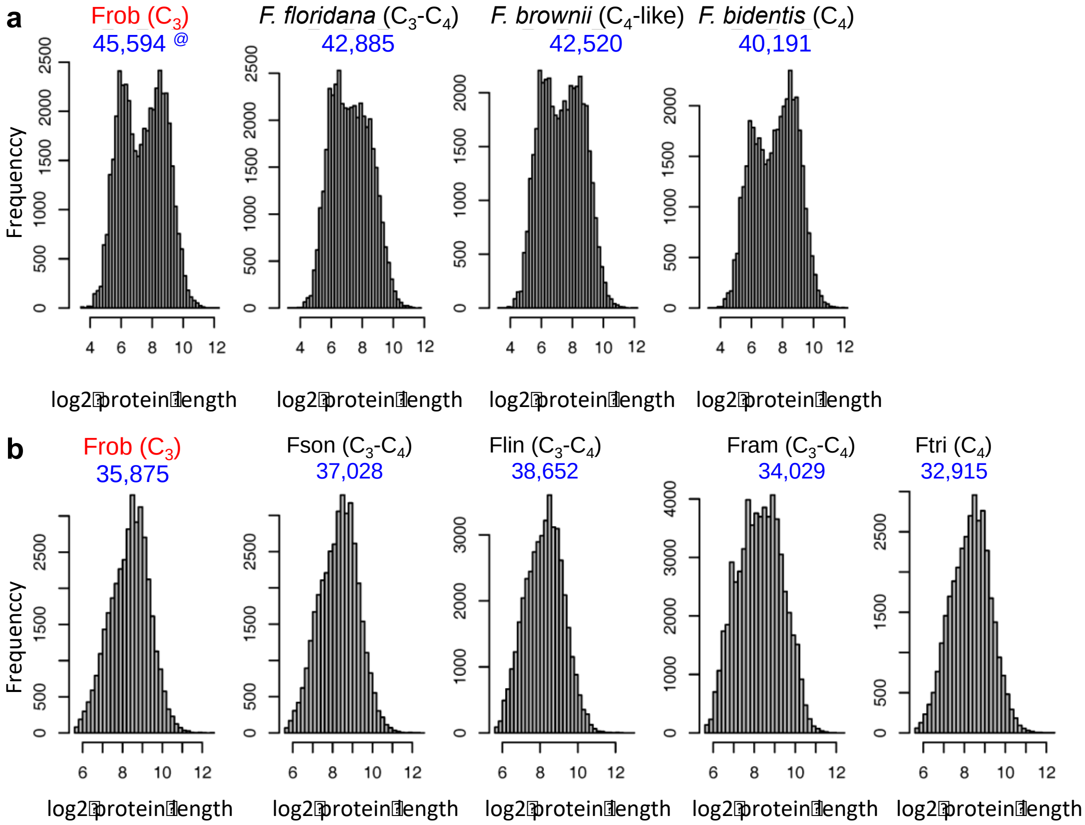
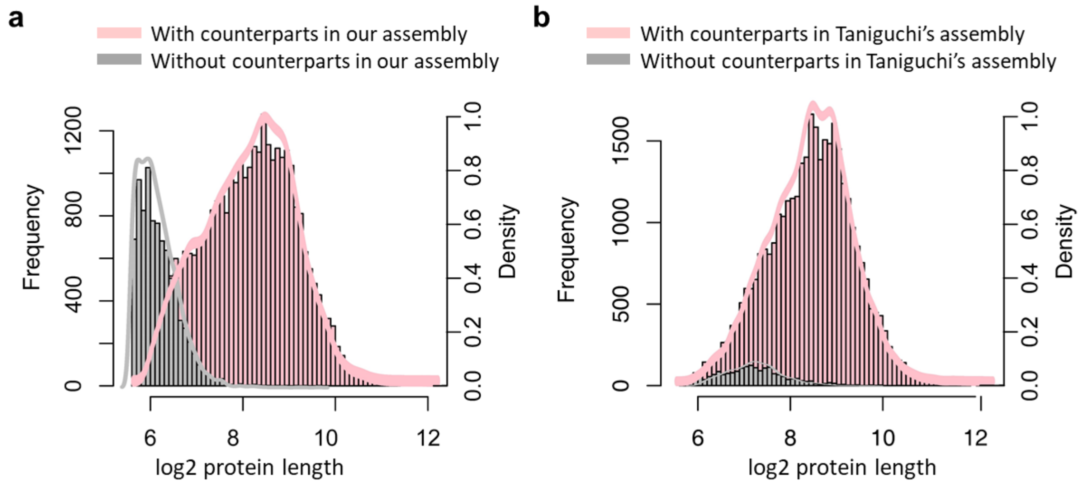


Fig. S 3**. Comparison of the protein length of predicted genes from our genome assembly as well as a published genome assembly**

(a) Histograms showing the distribution protein length of the four *Flaveria* species from Taniguchi's report [6]. (b) Histograms indicate the distribution protein length of the five *Flaveria* species from our study. The number of annotated protein-coding genes are indicated under the name of species in blue font. Frob is the only common species between the two studies. (Note: @ The number of protein-coding genes as reported originally in Taniguchi's assembly are 46138/43016/42802/40631 for Frob/*F. floridana*/*F. brownii*/ *F. bidentis* respectively. The number of protein-coding genes on the histograms are the number of genes with open reading frames, therefore, the numbers are slightly lower than those originally reported)

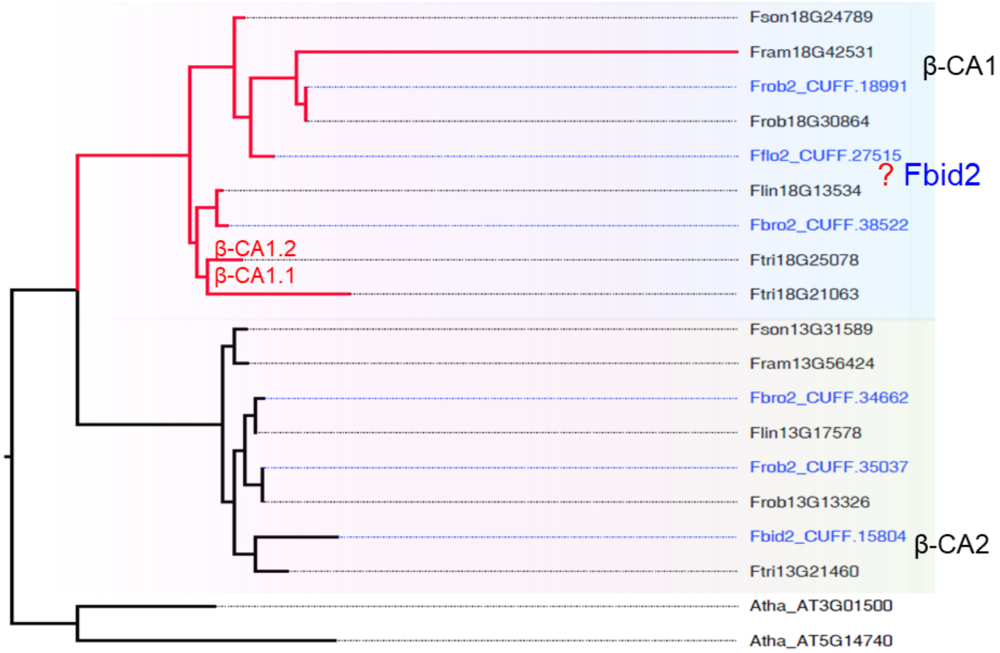


c

Frob	# Assembled gene	# Gene >=100 aa	# Gene >=100 and with ortholog in the other assembly	Percentage
Taniguchi's assembly	45,594$^{@}$	29,299	28,147	96.1%
Our assembly	35,875	34,388	33,123	96.3%

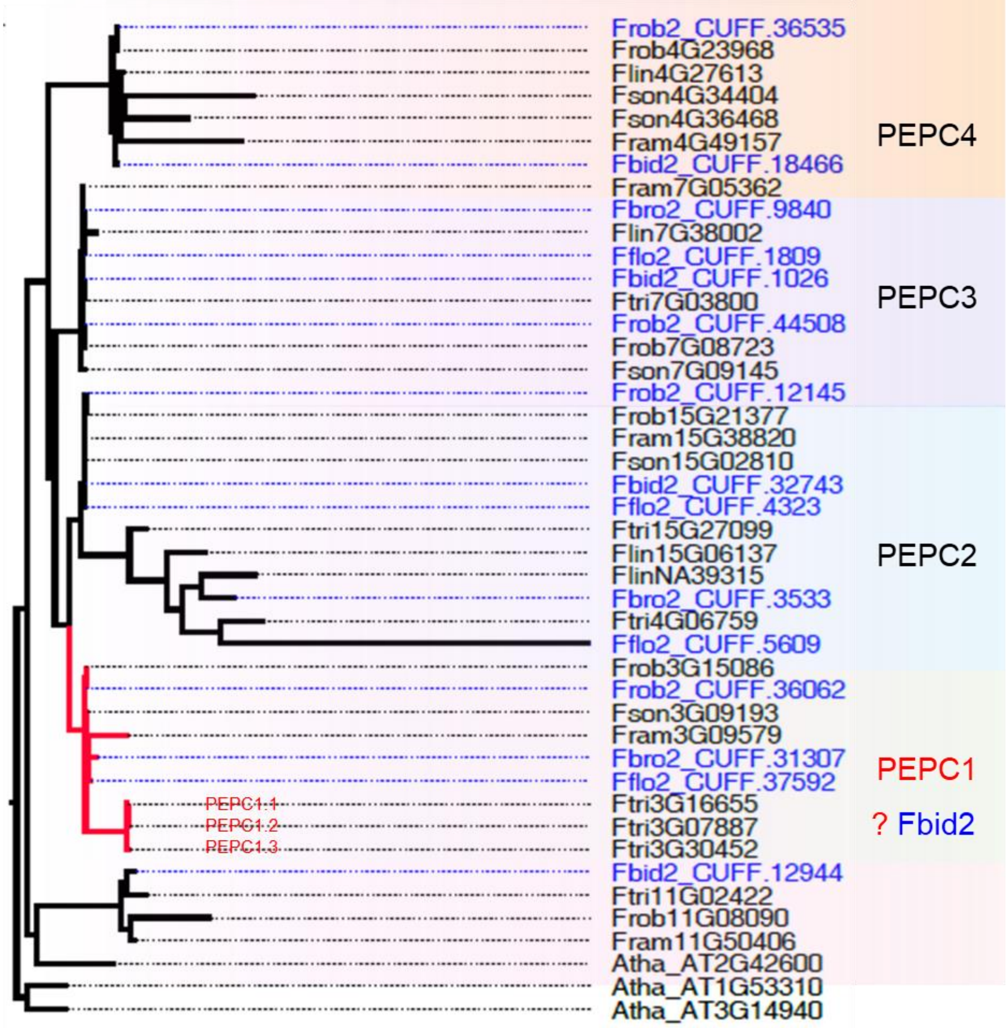
Fig. S 4. **Comparison of the annotated Frob protein from Taniguchi's assembly and our assembly**

Genes with counterparts in another assembly (either ours or Taniguchi's) were predicted. (a) The distribution of protein length of genes from Taniguchi's assembly with or without counterparts in our assembly are illustrated in pink and grey respectively. (b) The distribution of protein length of genes from our assembly with or without orthologs in Taniguchi's assembly are illustrated in pink and grey respectively. (c) Statistics of annotated protein coding genes and those with counterparts in the other assembly. (Note: @ The number of protein-coding gene reported originally in Taniguchi's assembly [6] is 46,138.)



● From published assemblies: Frob2 (C_3), Fflo2 (C_3-C_4), Fbro2 (C_4-like), Fbid2 (C_4)

● From our assemblies: Frob (C_3), Fson (C_3-C_4), Fram (C_3-C_4), Ftri (C_4)



● From published assemblies: Frob2 (C_3), Fflo2 (C_3-C_4), Fbro2 (C_4-like), Fbid2 (C_4)

● From our assemblies: Frob (C_3), Fson (C_3-C_4), Fram (C_3-C_4), Ftri (C_4)

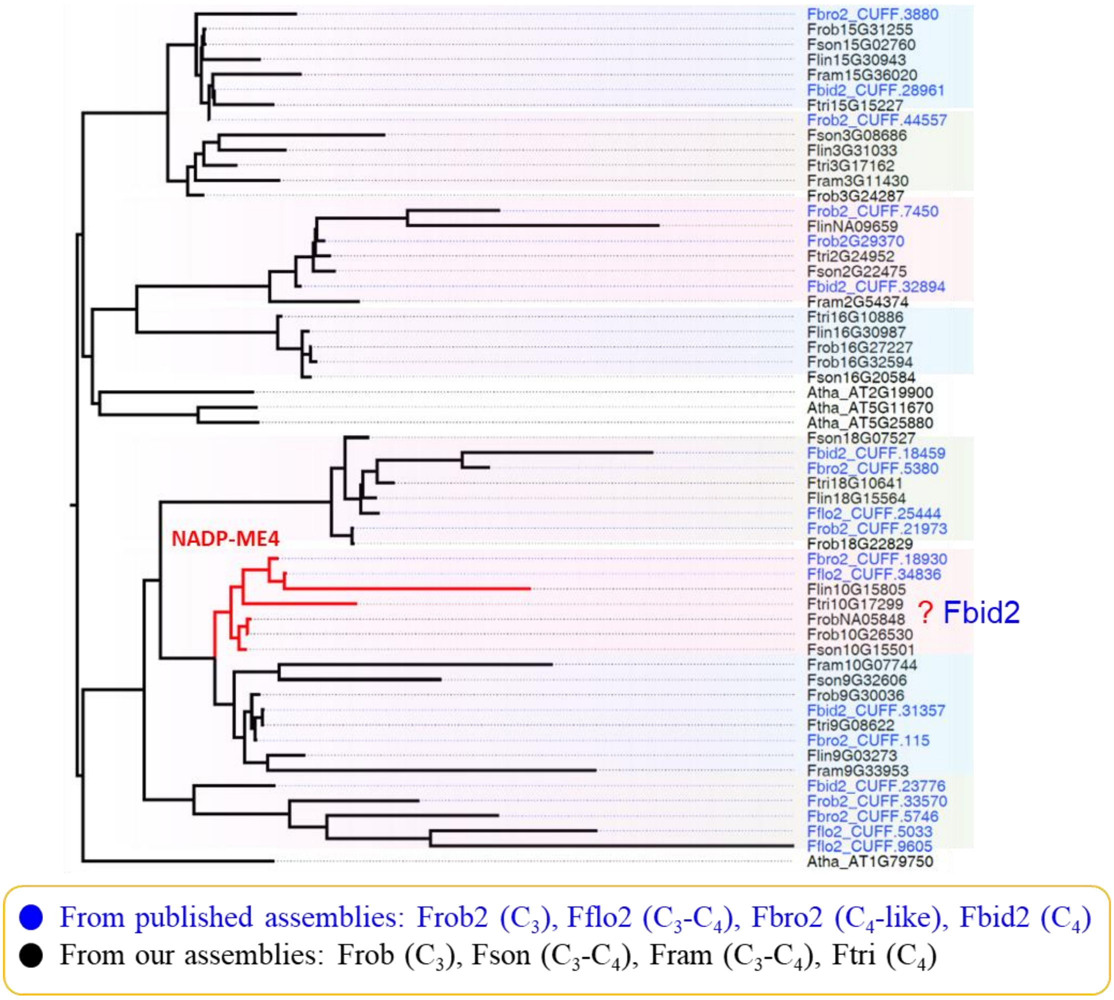


Fig. S 5. **The C_4 copies of CA1, PEPC1 and NADP-ME4 were not annotated in *F. bidentis* from published assembly**

The phylogenetic tree of *PEPC* was constructed based on protein sequences. Sequences of Frob2, Fflo2, Fbro2 and Fbid2 (labeled in blue font on the gene tree) are from Taniguchi's assembly, those of Frob, Fson, Fram and Ftri (labeled in black font) are from our assembly. *PEPC1* was determined as C_4 version, which was not found in Fbid2, which may be a result of uncompleted assembly in this species (Abbreviations: Frob: *F. robusta*, Fflo: F. *floridana*; Fbro: *F. brownii*; Fbid: *F. bidentis*, Fson: *F. sonorensis*; Fram: *F. ramosissima*, Ftri: F. trineriva).

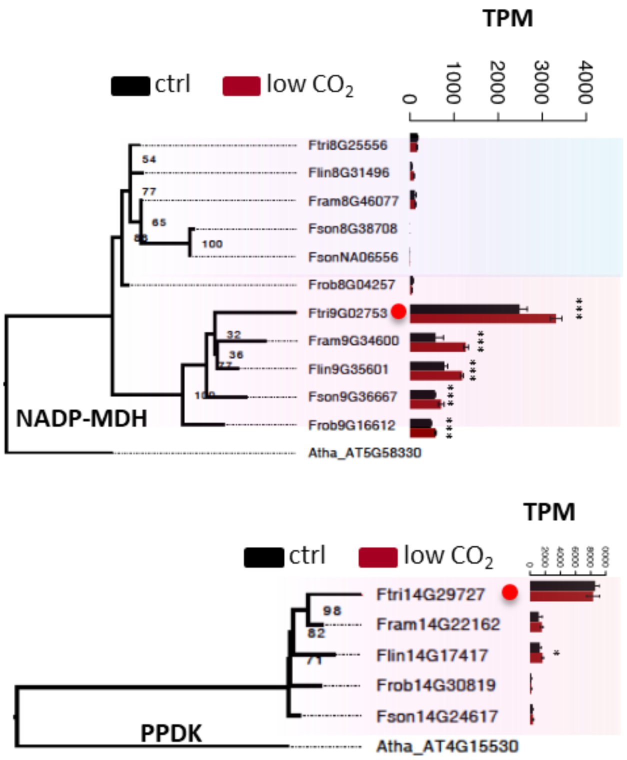
5. Determination of functional copies of C_4 genes

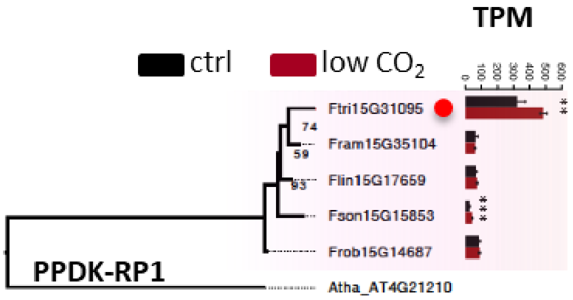
Methods

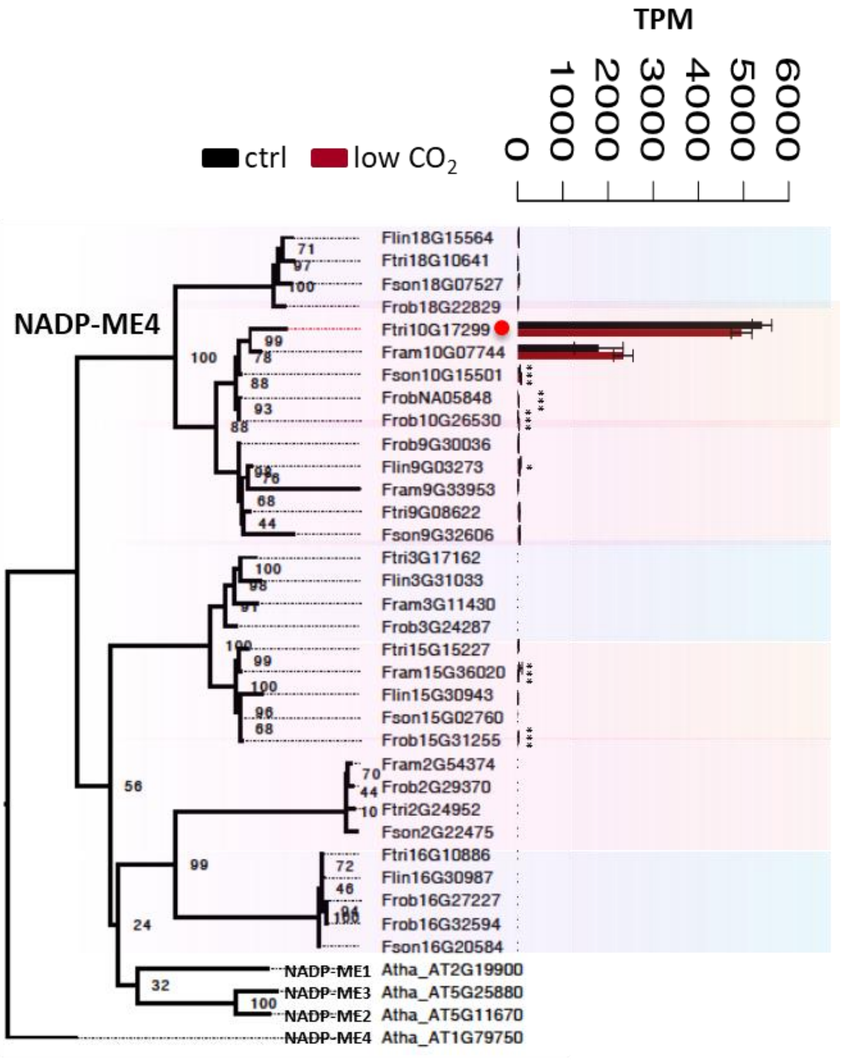
To determine the functional copies of C_4 genes, or C_4 version of C_4 genes, orthologous groups of C_4 genes were first characterized using Orthofinder (v2.3.11) [9] with default parameters. The C_4 versions were determined by combining gene phylogenetic tree and transcript abundances. To construct gene trees, protein sequences were aligned with MUSCLE (v3.8.31) [10] in default parameters, and Raxml (v7.9.3)[11] was applied to construct the gene tree based on the aligned protein sequences using the PROTGAMMAILG model (General Time Reversible amino acid substitution model with assumption that variations in sites follow gamma distribution). Transcript abundances were calculated based on RNA-seq data from a 4-week low CO_2 experiment.

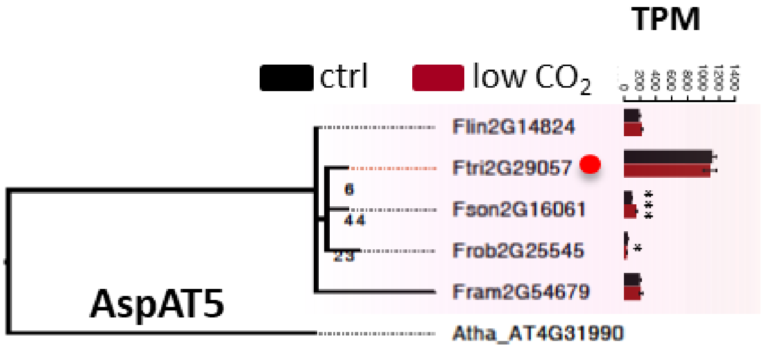
Results

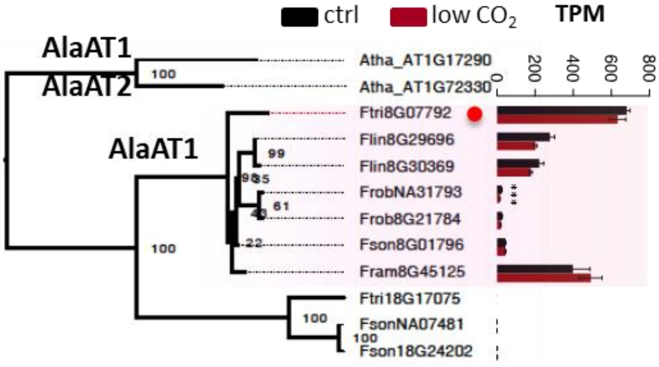
As most C_4 genes belong to gene family with multiple paralogs both across both C_3 and C_4 species, the C_4 version of C_4 genes was determined using the following two criteria: 1) showing higher transcript abundance in C_4 than in C_3 species, and 2) higher transcript abundance among its paralogous group. The C_4 versions of C_4 genes show higher transcript abundance in C_4 species than in other species, and usually are the highest paralog copy in C_4 species (Fig. S 6). For these determined C_4 version of C_4 genes, the evolutionary pattern of the C_4 version of C_4 genes in transcript and protein abundances were consistent (Fig. S 7).

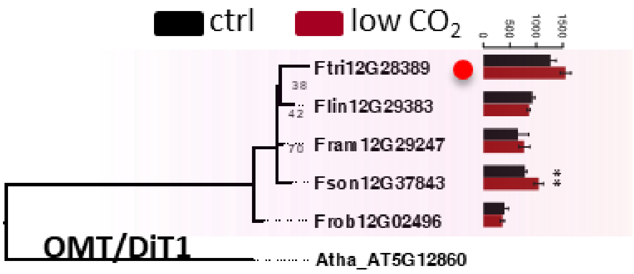


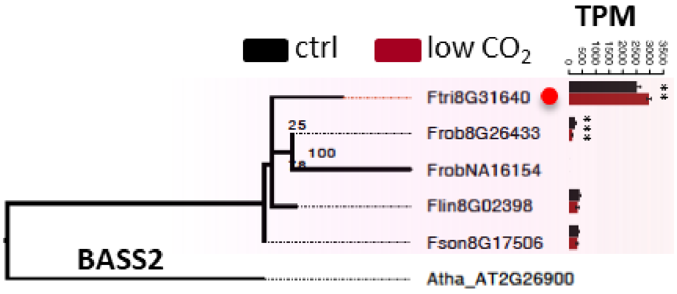

TPM
ctrl
low CO$_2$
0
100
200
300
400
500
600
Ftri15G31095
Fram15G35104
Flin15G17659
Fson15G15853
Frob15G14687
Atha_AT4G21210
74
59
93
**
***
PPDK-RP1

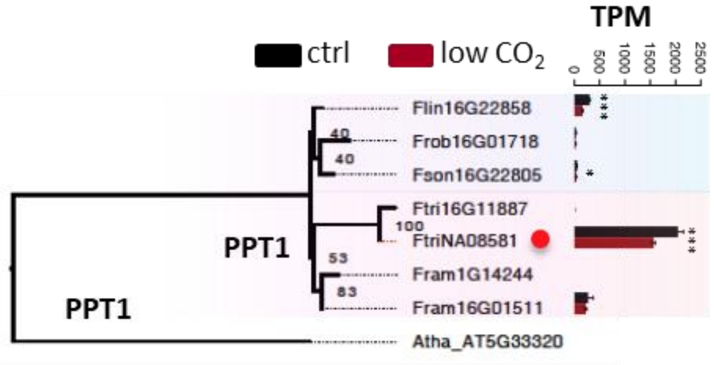

TPM
0
1000
2000
3000
4000
5000
6000
ctrl
low CO$_2$
NADP-ME4
Flin18G15564
Ftri18G10641
Fson18G07527
Frob18G22829
Ftri10G17299
Fram10G07744
Fson10G15501
FrobNA05848
Frob10G26530
Frob9G30036
Flin9G03273
Fram9G33953
Ftri9G08622
Fson9G32606
Ftri3G17162
Flin3G31033
Fram3G11430
Frob3G24287
Ftri15G15227
Fram15G36020
Flin15G30943
Fson15G02760
Frob15G31255
Fram2G54374
Frob2G29370
Ftri2G24952
Fson2G22475
Ftri16G10886
Flin16G30987
Frob16G27227
Frob16G32594
Fson16G20584
NADP-ME1 Atha_AT2G19900
NADP-ME3 Atha_AT5G25880
NADP-ME2 Atha_AT5G11670
NADP-ME4 Atha_AT1G79750
71
97
100
100
99
78
88
93
88
76
68
44
100
98
97
100
99
100
96
68
70
44
10
56
99
72
46
24
32
100
***
*


TPM
■ ctrl
■ low CO_2
0 200 400 600 800 1000 1200 1400
Flin2G14824
Ftri2G29057
6
44
Fson2G16061
***
23
Frob2G25545
*
Fram2G54679
AspAT5
Atha_AT4G31990

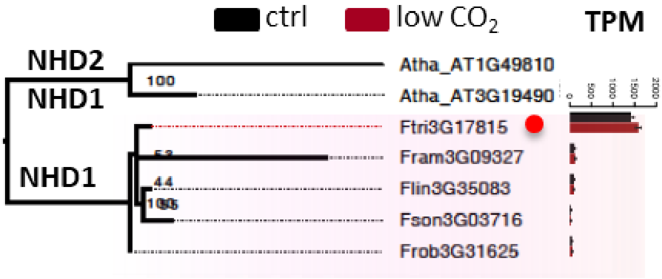

■ ctrl
■ low CO_2
TPM
AlaAT1
AlaAT2
100
Atha_AT1G17290
Atha_AT1G72330
0 200 400 600 800
Ftri8G07792
AlaAT1
Flin8G29696
99
Flin8G30369
98
35
FrobNA31793
***
61
48
Frob8G21784
22
Fson8G01796
100
Fram8G45125
Ftri18G17075
100
FsonNA07481
100
Fson18G24202


■ ctrl
■ low CO_2
0 500 1000 1500
Ftri12G28389
38
Flin12G29383
42
Fram12G29247
70
Fson12G37843
**
Frob12G02496
OMT/DiT1
Atha_AT5G12860

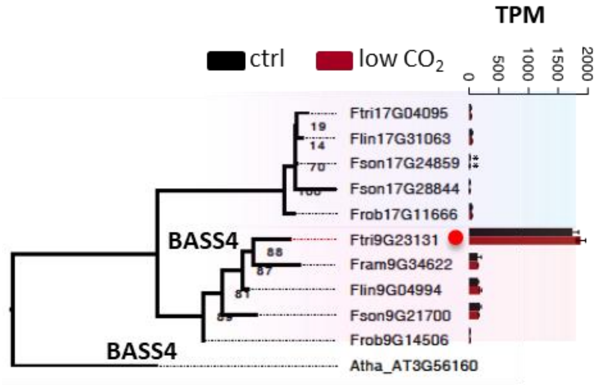

TPM
■ ctrl
■ low CO_2
0 500 1000 1500 2000 2500 3000 3500
Ftri8G31640
**
25
Frob8G26433
***
100
78
FrobNA16154
Flin8G02398
Fson8G17506
BASS2
Atha_AT2G26900



TPM
ctrl
low CO$_2$
0
500
1000
1500
2000
2500
Flin16G22858
Frob16G01718
Fson16G22805
Ftri16G11887
FtriNA08581
Fram1G14244
Fram16G01511
Atha_AT5G33320
40
40
100
53
83
PPT1
PPT1
***
*
***



ctrl
low CO$_2$
TPM
NHD2
NHD1
NHD1
100
Atha_AT1G49810
Atha_AT3G19490
Ftri3G17815
Fram3G09327
Flin3G35083
Fson3G03716
Frob3G31625
44
0
500
1000
1500
2000



TPM
ctrl
low CO$_2$
0
500
1000
1500
2000
Ftri17G04095
Flin17G31063
Fson17G24859
Fson17G28844
Frob17G11666
Ftri9G23131
Fram9G34622
Flin9G04994
Fson9G21700
Frob9G14506
Atha_AT3G56160
19
14
70
88
87
81
BASS4
BASS4

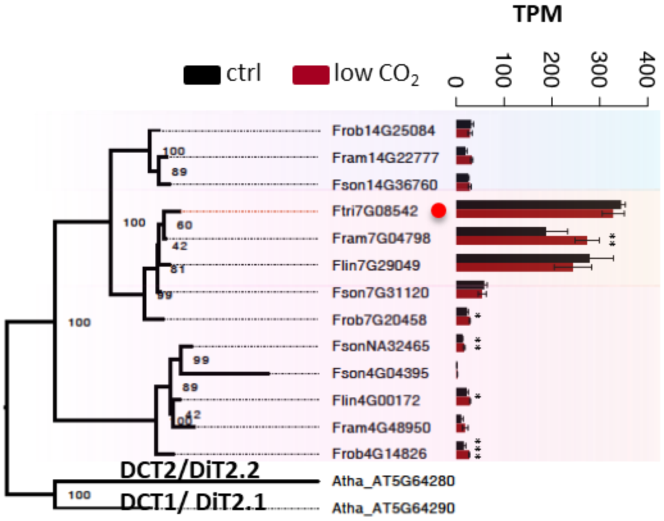


Fig. S 6. **Determining the** C_4 **version of** C_4 **genes combining gene tree and transcript abundances**

Gene tree and transcript abundances of each C_4 genes from 4-week low CO_2 experiment are shown. Genome-wide orthologous groups were predicted using Orthofinder. Gene trees were constructed using protein sequences from Raxml. Bootstraps were calculated from 100 independent trees. Gene expression in transcript per million mapped reads (TPM) is shown, where ctrl represents TPM of plant growing in normal CO_2 conditions (380 PPM) and low CO_2 represents TPM of plant growing in low CO_2 conditions (100 PPM) (n=3). Standard deviations are from three biological replicates. The predicted functional versions of C_4 genes are indicated with red circles.

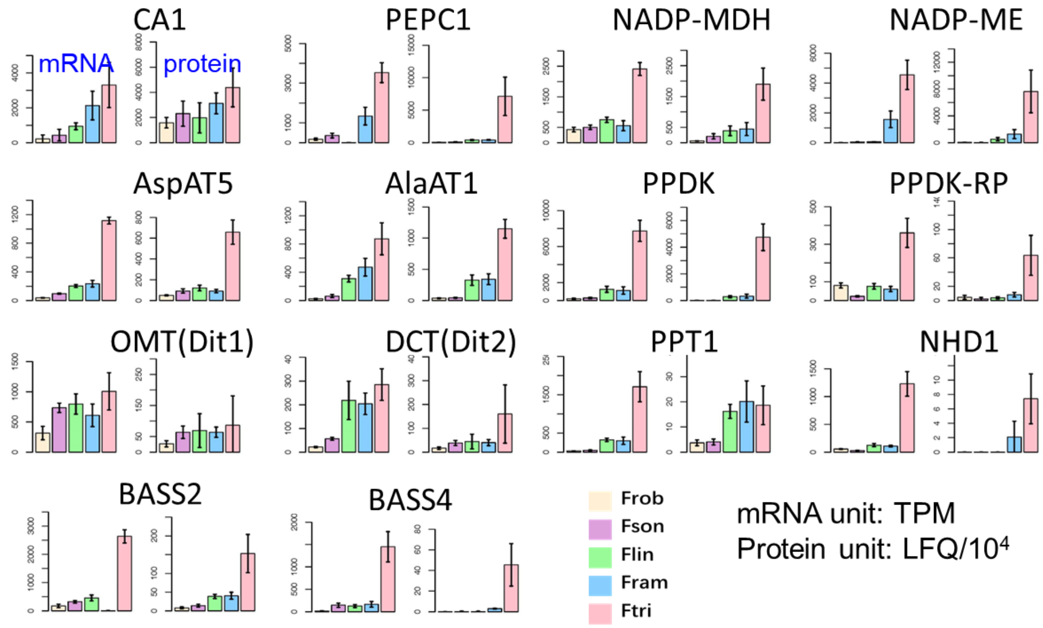


Fig. S 7**. C$_4$ genes show a consistent evolutionary profile in transcript and protein abundances across the five *Flaveria* species**

Bar plots illustrate the transcript abundances (left) and protein abundances (right) of C$_4$ enzyme and transporters from the five *Flaveria* species. The Y-axis represents transcript per million mapped reads (TPM) for transcript abundances and label free quantification (LFQ)/10^4 for protein abundance (Abbreviations: TPM: transcript per million mapped reads, LFQ: label free quantification, *CA1*, *carbonic anhydrase 1*; *PEPC1*: *phosphoenolpyruvate carboxylase 1*; *NADP-MDH*: *NADP-dependent malate dehydrogenase*; *NADP-ME4*: *NADP-dependent malic enzyme 4*; *AspAT*: *Aspartate amino acid transport*; *AlaAT*: *alanine aminotransferase*; *PPDK*: *pyruvate orthophosphate dikinase*; *PPDK-RP*: *PPDK regulatory protein*; *OMT*: *oxaloacetate/malate transporter or dicarboxylate transporter 1* (*DiT1*); *DCT*: *dicarboxylate transport 2.1* (or *DiT2.1*); *PPT1*: *phosphate/phosphoenolpyruvate translocator 1*; *NHD1*: *sodium: hydrogen antiporter 1*;*BASS2*: *bile acid sodium symporter 2*; *BASS4*: *bile acid sodium symporter 4*; *PEPC kinase* (*PEPC-k*) was not included as the C$_4$ version of *PEPC-k* was not detected in Fram;)

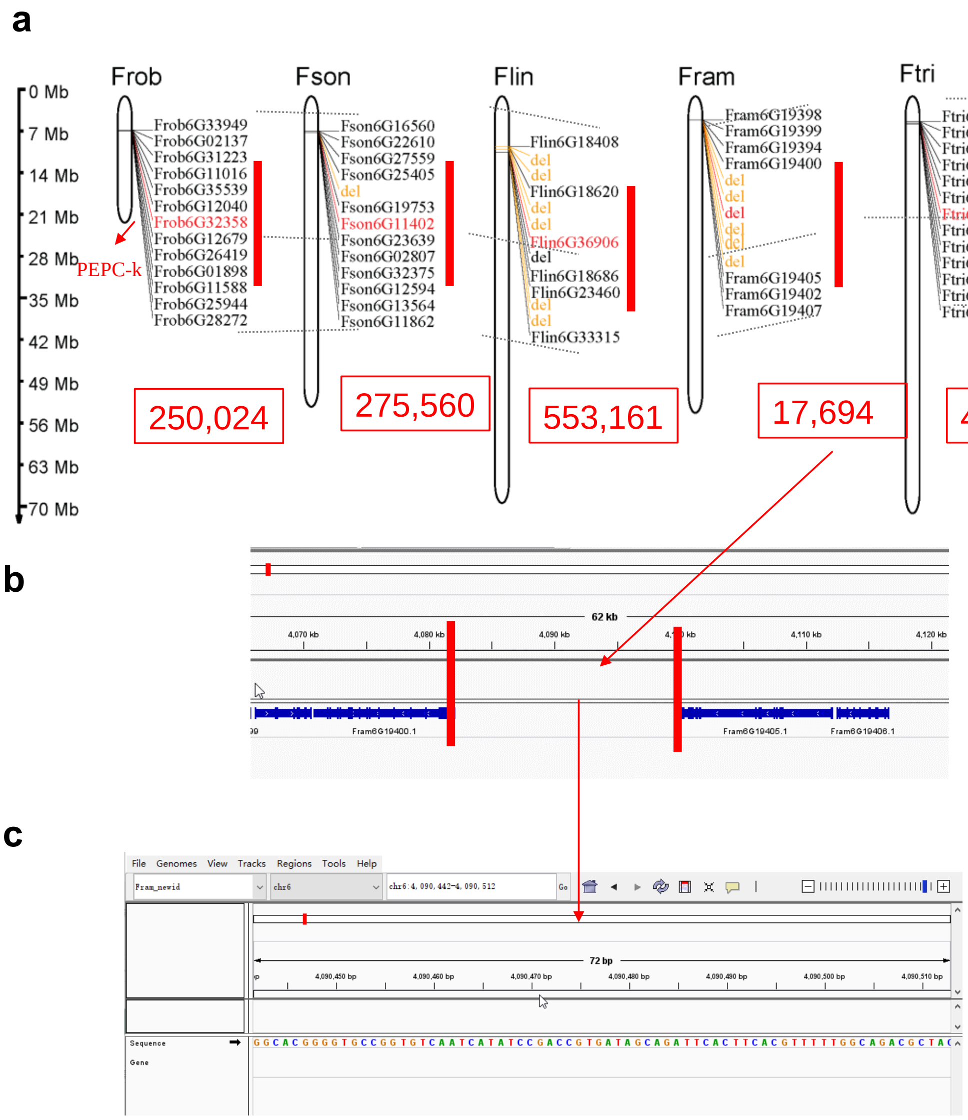
6. C$_4$ version of *PEPC-k* was absent in Fram plant sequenced in this study

Methods

We found that the C$_4$ version *PEPC-k* was absent in Fram (Fig. 1a). We checked whether this was due to an error in genome assembly. We investigated the presence and absence of adjacent genes to *PEPC-k* in all five *Flaveria* species. Specifically, we analyzed six genes upstream of *PEPC-k* and six genes downstream of *PEPC-k* according to the chromosome location of Frob (13 genes in total). We then checked whether the orthologous genes were present in the other four *Flaveria* species.

Results

We found that the counterparts of the 13 genes were also present in the C_4 species Ftri, while, one of the orthologous genes was absent in Fson. Notably, 7 genes in tandem were absent in the Flin genome, and 6 genes in tandem were absent in the Fram genome, including *PEPC-k* (Fig. S 8 a). This suggests that the genome segments containing *PEPC-k* were dynamic during the evolution in the genus of *Flaveria*. We then checked whether the deletion in Fram was due to an assembly gap. We found that the sequence is not “N”, suggesting that this may be not a gap (Fig. S 8 b and c). Therefore, our results suggested that the Fram plant sequenced here had a deletion of the C_4 version of *PEPC-k*.



G G C A C G G G G T G C C G G T G T C A A T C A T A T C C G A C C G T G A T A G C A G A T T C A C T T C A C G T T T T T G G C A G A C G C T A C

Fig. S 8. **Syntenic chromosome region containing *PEPC-k* across the five *Flaveria* species** (a) The syntenic chromosome region containing *PEPC-k*. Six genes from upstream and downstream of *PEPC-k* according to Ftri are also shown. *PEPC-k* is labeled in red font, a deletion is represented by "del" in orange. The number in the red frame indicates the length of the chromosome segment. (b) IGV indicates the chromosome region of the deletion of the six genes (between Fram6G19400 and Fram6G19405, including *PEPC-k*) in Fram. (c) The nucleotide sequences between Fram6G19400 and Fram6G19405. Note that the sequence is not "N", suggesting that the deletion is not due to an assembly gap.

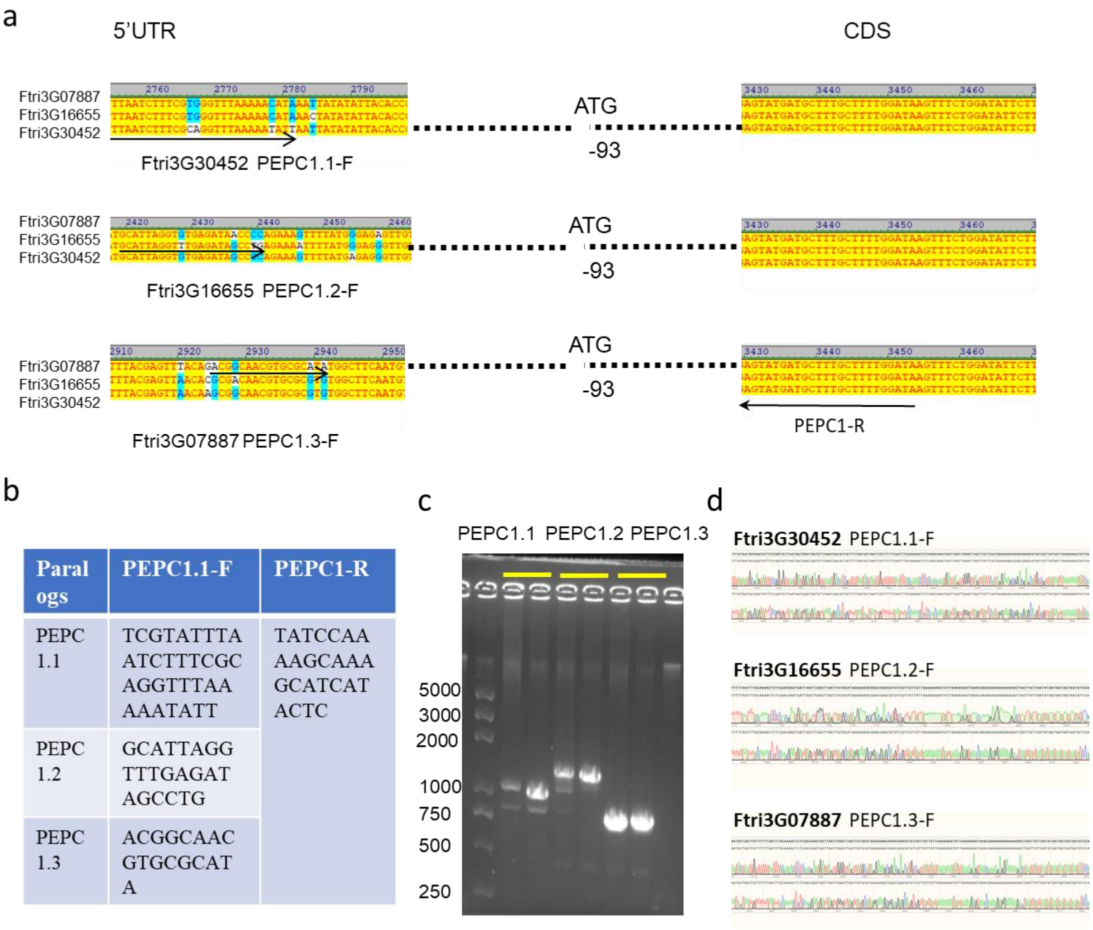
7. Verification of three copies of PEPC1 in the C_4 species Ftri

Methods

To verity the presence of three copies of *PEPC1* in Ftri, genomic DNA was isolated from young leaves of one-month old plants. DNA isolation was performed as described previously [12]. Forward primers were designed to be complementary to the non-conserved region of the 5' UTR of the three *PEPC1* copies respectively, and the reverse primer was designed to be complementary the conserved region of the coding sequence. The forward primers were: TCGTATTTAATCTTTCGCAGGTTTAAAAATATT for Ftri3G07887 (*PEPC1.1*), GCATTAGGTTTGAGATAGCCTG for Ftri3G16655 (*PEPC1.2*), and ACGGCAACGTGCGCATA for Ftri3G30452 (*PEPC1.3*). The reverse primer was identical for three loci: TATCCAAAAGCAAAGCATCATACTC.

Results

The three copies of *PEPC1* in the C_4 species Ftri were verified by PCR. (Fig. S 9).



Paralogs	PEPC1.1-F	PEPC1-R
PEPC 1.1	TCGTATTTAATCTTTCGCAGGTTTAAAAATATT	TATCCAAAAGCAAAGCATCATACTC
PEPC 1.2	GCATTAGGTTTGAGATAGCCTG	
PEPC 1.3	ACGGCAACGTGCGCATA	

Fig. S 9**. Verification of three *PEPC1* paralogs in Ftri using PCR**

(a) Schematic of the designed forward and reverse primers of there *PEPC1*s in Ftri. The forward primers are complementary to the 5'UTR where the sequences are different between the three *PEPC1*s. The reverse primer is complementary to the conserved protein coding region. Primer sequences for each *PEPC1* paralog are shown in (b), gel results and Sanger sequencing are shown in (c) and (d).

8. Investigation of tandem duplications of C_4 genes in other C_4 species

Methods

To investigate whether C_4 genes show tandem duplication in other C_4 species, we expanded the evolution of C_4 genes to more species. In total, two dicotyledonous C_4 species, *i.e.*, Ftri and *Gynandropsis gynandra* (Ggyn), and six monocotyledonous C_4 species, *i.e.*, *Zea mays* (V5), *Setaris viridis* (V2.1), *Panicum Hallii* (V2.1), *Sorghum biocolor* (v3.0.1), and *Miscanthus lutarioriparius* (V1) were included in the comparisons. Additionaly, six dicotyledonous C_3 species and two monocotyledonous C_3 species, as well as one algal species (*Chlamydomonas reinhardtii*) were included. The protein sequences of those species were downloaded from public databases as mentioned in the Methods of the manuscript. Orthologous genes were predicted using Orthofinder (v2.3.11) [9] with default parameters, orthologous protein sequences were aligned by MUSCLE (v3.8.31) [10] with default parameters, andgene trees were constructed using Raxml (v7.9.3) [11] based on alignments of protein sequences with the PROTGAMMAILG model (General Time Reversible amino acid substitution model with assumption that variations in sites follow gamma distribution).

To investigate whether the same orthologous genes were co-opted in dicotyledonous and monocotyledonous C_4 species, we quantified transcript abundances of genes for all C_4 species (Fig. S 10 a) and several C_3 species, including *Oryza sativa*, *Arabidopsis thaliana*, *Tarenaya hassleriana*, and the algae species *Chlamydomonas reinhardtii*. The RNA-seq data used here are mentioned in Supplementary Note 14.

Results

Our results indicated that tandem duplication of C_4 version of *CA*, *PEPC*, and *PEPC-k* were not universal but rather, species-specific (Fig. S 10). Notably, we found that the same orthologous genes were recruited for C_4 photosynthesis across different C_4 species. Remarkably, the counterparts of recruited C_4 genes already showed high transcript abundances in C_3 species.

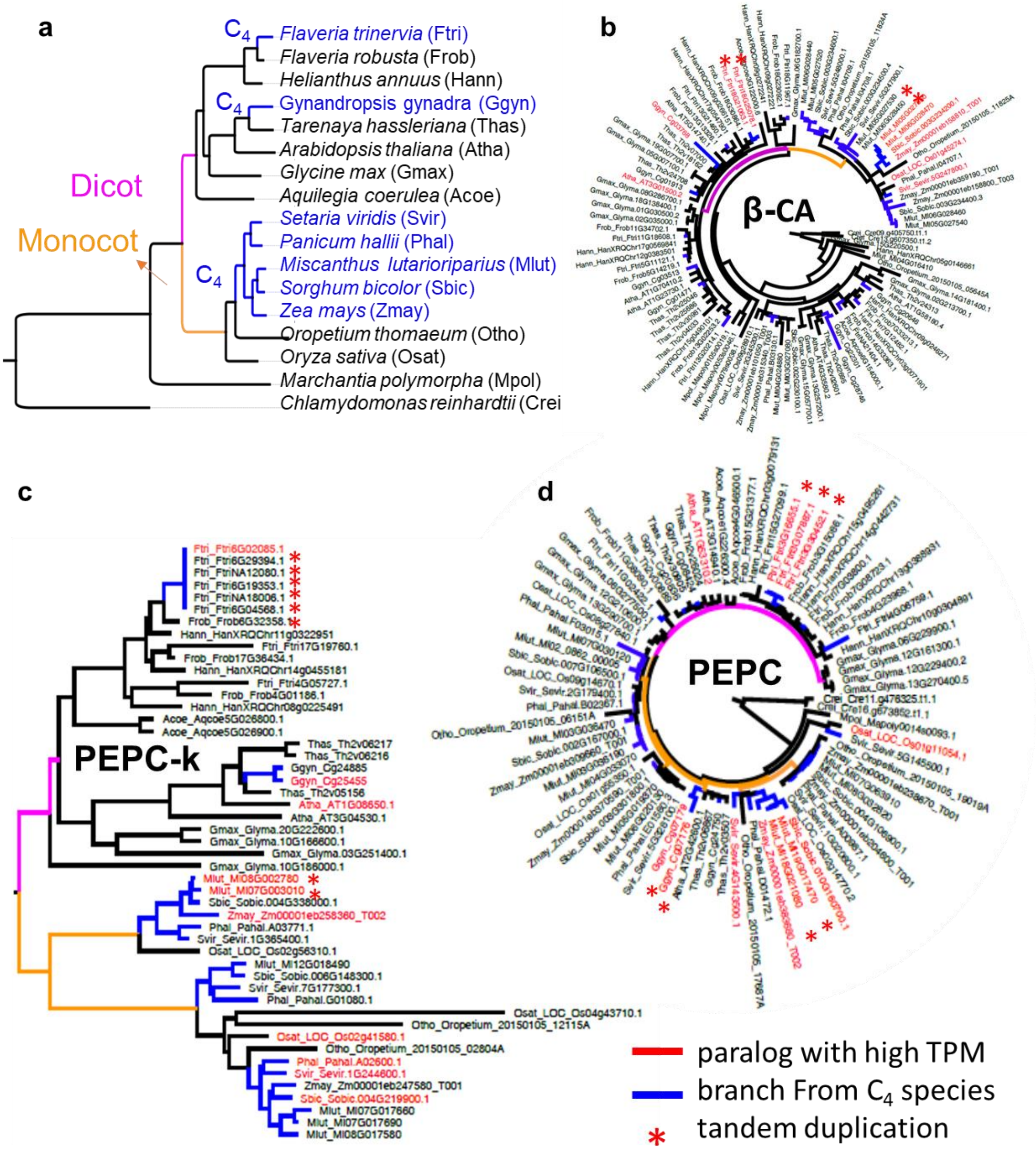

a
C$_4$
Flaveria trinervia (Ftri)
Flaveria robusta (Frob)
Helianthus annuus (Hann)
C$_4$
Gynandropsis gynadra (Ggyn)
Tarenaya hassleriana (Thas)
Arabidopsis thaliana (Atha)
Glycine max (Gmax)
Aquilegia coerulea (Acoe)
Dicot
Monocot
C$_4$
Setaria viridis (Svir)
Panicum hallii (Phal)
Miscanthus lutarioriparius (Mlut)
Sorghum bicolor (Sbic)
Zea mays (Zmay)
Oropetium thomaeum (Otho)
Oryza sativa (Osat)
Marchantia polymorpha (Mpol)
Chlamydomonas reinhardtii (Crei
b
β-CA
c
PEPC-k
d
PEPC
paralog with high TPM
branch From C$_4$ species
* tandem duplication

Fig. S 10. **Tandem duplications of** C$_4$ **version of** ***CA*****,** ***PEPC*** **and** ***PEPC-k*** **are not universal but species-specific**

(a) Phylogenetic tree of species to investigate tandem duplication of C$_4$ genes. Phylogenetic relationships were inferred from the Phytozome website (https://phytozome-next.jgi.doe.gov). C$_4$ species are labeled in blue. Abbreviations of each species using in the following gene tress are shown in parentheses after each species. The gene trees of *CA* (b), *PEPC-k* (c), and *PEPC* (d) are shown, where the main branch of monocotyledonous plants are shown in orange with dicotyledonous plants in purple. Branches labeled in blue are C$_4$ species, the gene with the highest TPM is labeled in red font. The format of the gene on the gene tree is as follows: “abbreviation of species name” pluses gene id jointed with “_”.

9. Analysis of transposable elements and their effects on duplicated Ftri*PEPC1*

Methods

To predict transposable elements (TEs), entire genome sequences of the five *Flaveria* species were searched for repetitive sequences individually. A *de novo* repeat sequence library was obtained using RepeatModeler (RepeatModeler-Open-1.0.5) with the following parameters: RepeatModeler -database database_name -engine ncbi -pa [int]. We then used RepatMasker (RepeatMasker-Open-4.1.0) to search for similar TEs against the de novo library with the following parameters: RepeatMasker genome.fa -lib de_novo_library -nolow -no_is -q -engine rmblast -pa [int] –norna. Intact long terminal repeat retrotransposons (LTR-RTs) were identified using LTR_FINDER (v1.07) [13] and LTRharvest (v1.5.10) [14] with default parameters. LTR_Retriever (v2.9.0) [15] was then used to merge the above results with the following parameters: LTR_retriever -genome genome.fa -inharvest species.harvest.scn -infinder species.finder.scn –nonTGCA species.harvest.nonTGCA.scn. The insertion time of intact LTR-RT was obtained from LTR-Retriever analysis.

Results

Transposable elements (TEs) showed the highest abundance in the C_4 species, where they accounted for 82% of the total genome, followed by C_3-C_4 species (from 65.6% to 71.8%), while the percentage in the C_3 species was 47.1% (Table S 5). In all five species, long terminal repeat retrotransposons (LTR-RTs) comprised the majority of the TEs, accounting for an average of 76% of the total TEs (from 42% to 91%). (Table S 6 and Table S 7). Flin and Fram had more evolutionary recent LTR-RTs than the other species (Fig. S 11). Compared to C_3 and intermediate species, C_4 species Ftri had a higher proportion of TEs at the whole genome scale.

We found that abundant retrotransposons were on the chromosomal regions between tandem duplicated *FtriCA1*, and the region between tandem duplicated *FtriPEPC1* and the region between tandem duplicated *FtriPEPC-K1* (Fig. S 12). In comparison, few retrotransposons were observed in the chromosome near the region containing *FtriPPDK*, which was a single copy gene.

Reports have showed that gene duplications could be mediated by retrotransposons through retroposition[16] [17] [18] [19] [20] [21]. To test whether the observed C_4 gene duplications were related to retrotransposons in Ftri, we took a close inspection of the evolution and sequences of Ftri*PEPC1*s. Among the three Ftri*PEPC1s*, Ftri*PEPC1.1* was predicted to be the ancestral copy, as the mesophyll expression module 1 (MEM1) on the promoter of Ftri*PEPC1.1* was conserved with that of *PEPC1* from other four *Flaveria* species, i.e., Frob, Fson, Fram and Flin, whereas, the MEM1 of Ftri*PEPC1.2* and Ftri*PEPC1.3* showed a deletion of 109 bps (Fig. S 13). Except for coding region, sequences from ~2500bp upstream and ~2000 bp downstream of the coding sequences of Ftri*PEPC1s* were also conserved among the three Ftri*PEPC1*s (Fig. S 14 b). Through close inspection of the sequences near the conserved region, we observed a 9-bp inverted repeat sequences, *i.e.*, 5'-AAAATAAAG-3'. Besides, a 4-bp motif, *i.e*., 5'-TTTT-3', immediately flanked the invert repeats (Fig. S 14 c).

To further test whether retrotransposons were related to duplications of DNA segments, we examined chromosomal regions containing duplicated copies of *CA1*, *PEPC1* and *PEPC-k1* in Ftri. We found that these chromosomal regions contain many duplicated sequences (Fig. S 15). Taken together, our results suggest that the gene duplications of C_4 genes might be mediated by retrotransposons.

Table S 5. **Compositions and proportions of transposable elements across the five *Flaveria* species**

	Frob	Fson	Flin	Fram	Ftri
	(% of genome)	(% of genome)	(% of genome)	(% of genome)	(% of genome)
Total Repeat Fractions	47.09	70.26	71.81	65.56	82.06
ClassⅠ :Retrotransposon	20.41	57.27	59.52	56.29	75.97
LTR Retrotransposon					
Ty1/Copia	12.69	34.1	27.32	38.51	56.6
Ty3/Gypsy	5.97	22.21	29.44	16.81	18.48
Other	1.18	0.72	0.66	0.74	0.32
NON-LTR Retrotransposon					
SINE	0.03	0.02	0.01	0.01	0
LINE	0.54	0.22	2.09	0.22	0.57
Class Ⅱ :DNA transposon	6.23	4.73	5.1	4.32	1.84
CACTA	0.82	0.71	1	0.82	0.37
Mutator	0.45	0.61	1.48	0.39	0.23
PIF-Harbinger	0.27	0.31	0.27	0.11	0.1
RC/Helitron	0.4	0.64	0.25	0.2	0.11
hAT	2.21	1.62	0.77	0.88	0.6
Other	2.08	0.84	1.33	1.92	0.43
Tandem Repeat	0.38	0.86	1.86	0.45	0.31
Unknown	20.07	7.4	5.33	4.5	3.94

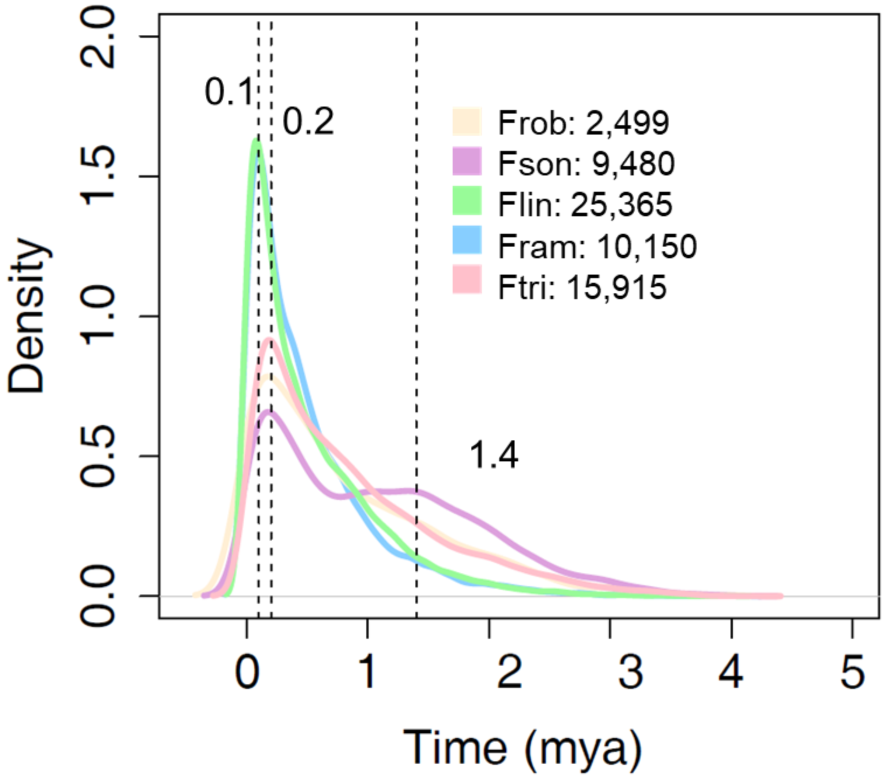


Fig. S 11**. The burst time of LTR-RTs estimated based on intact LTR-RTs**

The distribution of burst time of the contact LTR-RTs of the five *Flaveria* species are shown. The numbers of contact LTR-RTs for each species are indicated. (Abbreviations of species:

Frob, *F. robusta*, Fson, *F. sonorensis*, Flin: *F. linearis*, Fram: *F. ramosissma*, Ftri: *F. trinervia*.)

Table S 6**. The length of different types of transposon elements**

Species	**Total TE length (Mbp)**	**Total LTR-RT length (Mbp)**	**Total intact LTR-RT length (Mbp)**	**% Intact LTR-TR to TE**	**% Intact LTR-RT to LTR-RT**
Frob (C_3**)**	259.73	109.42	14.68	5.65%	13.42%
Fson (C_3-C_4**)**	872.90	708.48	80.89	9.27%	11.42%
Flin (C_3-C_4**)**	1188.64	950.59	236.96	19.94%	24.93%
Fram (C_3-C_4**)**	830.21	709.81	101.38	12.21%	14.28%
Ftri (C_4**)**	1480.09	1357.84	168.89	11.41%	12.44%

Note: LTR-RT: long terminal repeat retrotransposons, TE: transposon elements

Table S 7**. The different types of TEs**

Species	**TE**	**LTR-RT**	**Intact LTR-RT**	**% Intact LTR-RT to TE**	**% Intact LTR-RT to LTR-RT**
Frob (C_3**)**	571,290	129,171	2,499	0.44%	1.93%
Fson (C_3-C_4**)**	985,595	532,042	9,480	0.96%	1.78%
Flin (C_3-C_4**)**	1,298,249	767,238	25,365	1.95%	3.31%
Fram (C_3-C_4**)**	747,356	371,205	10,150	1.36%	2.73%
Ftri (C_4**)**	1,045,859	687,601	15,915	1.52%	2.31%

Note: LTR-RT: long terminal repeat retrotransposons, TE: transposon elements

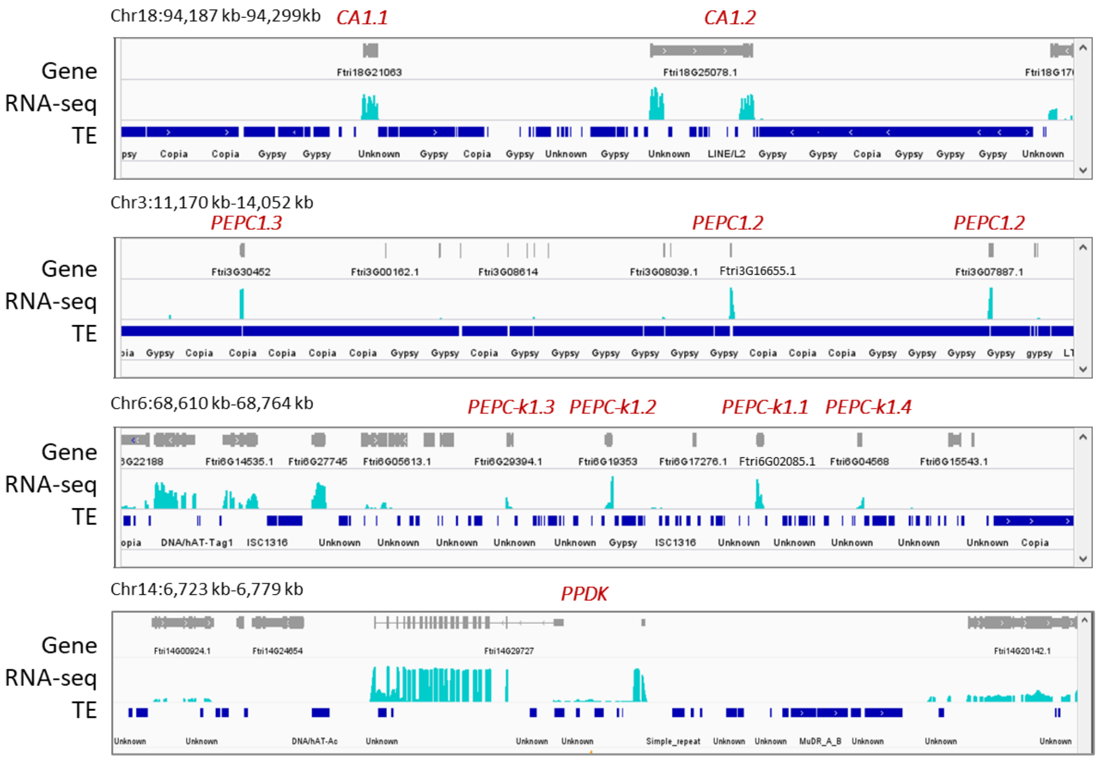
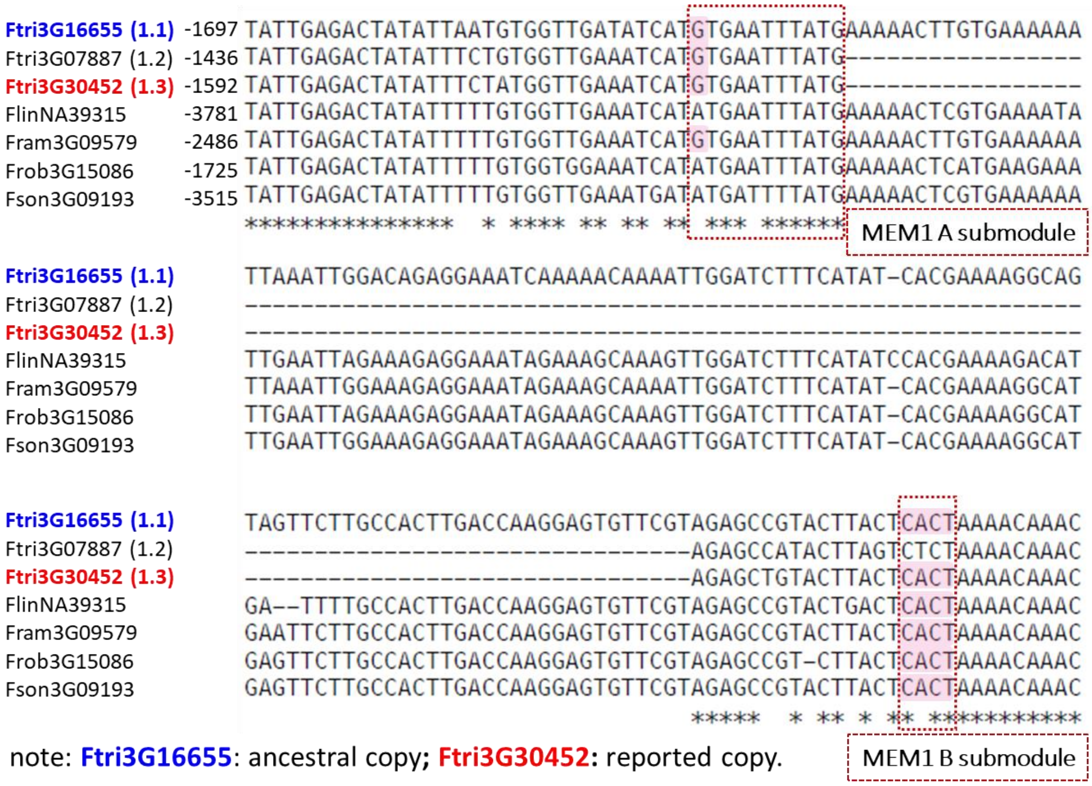

Chr18:94,187 kb-94,299kb
CA1.1
CA1.2
Gene
RNA-seq
TE
Ftri18G21063
Ftri18G25078.1
Ftri18G17
psy Copia Copia Gypsy Gypsy Unknown Gypsy Copia Gypsy Unknown Gypsy Unknown LINE/L2 Gypsy Gypsy Copia Gypsy Gypsy Gypsy Unknown
Chr3:11,170 kb-14,052 kb
PEPC1.3
PEPC1.2
PEPC1.2
Gene
RNA-seq
TE
Ftri3G30452
Ftri3G00162.1
Ftri3G08614
Ftri3G08039.1
Ftri3G16655.1
Ftri3G07887.1
pia Gypsy Copia Copia Copia Copia Copia Gypsy Gypsy Copia Gypsy Gypsy Gypsy Gypsy Gypsy Gypsy Copia Copia Copia Gypsy Gypsy Gypsy Gypsy gypsy L1
Chr6:68,610 kb-68,764 kb
PEPC-k1.3
PEPC-k1.2
PEPC-k1.1
PEPC-k1.4
Gene
RNA-seq
TE
6G22188
Ftri6G14535.1
Ftri6G27745
Ftri6G05613.1
Ftri6G29394.1
Ftri6G19353
Ftri6G17276.1
Ftri6G02085.1
Ftri6G04568
Ftri6G15543.1
opia DNA/hAT-Tag1 ISC1316 Unknown Unknown Unknown Unknown Unknown Gypsy ISC1316 Unknown Unknown Unknown Unknown Unknown Copia
Chr14:6,723 kb-6,779 kb
PPDK
Gene
RNA-seq
TE
Ftri14G00924.1
Ftri14G24654
Ftri14G29727
Ftri14G20142.1
Unknown Unknown DNA/hAT-Ac Unknown Unknown Unknown Simple_repeat Unknown Unknown MuDR_A_B Unknown Unknown Unknown

Fig. S 12**. Retrotransposons may be related to tandem duplication of *CA1*, *PEPC1* and *PEPC-k1* in Ftri.**

Integrative genomics viewer (IGV) indicating the genes, transcript abundances and predicted TEs on the chromosomal regions between tandem-duplicated *CA1*, *PEPC1* and *PEPC-k1*. Transcript abundances are shown in transcript per million mapped reads (TPM). The chromosome region containing *PPDK* is also illustrated (bottom panel). (Abbreviations: *CA1*, *carbonic anhydrase 1*; *PEPC1*, *phosphoenolpyruvate carboxylase 1*; *PEPC-K*, *PEPC kinase*; *PPDK*, *pyruvate/orthophosphate dikinase*.)



```
Ftri3G16655 (1.1) -1697 TATTGAGACTATATTAATGTGGTTGATATCATGTGAATTTATGAAAAACTTGTGAAAAAA
Ftri3G07887 (1.2) -1436 TATTGAGACTATATTTCTGTGGTTGAAATCATGTGAATTTATG-----------------
Ftri3G30452 (1.3) -1592 TATTGAGACTATATTTCTATGGTTGAAATCATGTGAATTTATG-----------------
FlinNA39315       -3781 TATTGAGACTATATTTTTGTGGTTGAAATCATATGAATTTATGAAAAACTCGTGAAAATA
Fram3G09579       -2486 TATTGAGACTATATTTTTGTGGTTGAAATCATGTGAATTTATGAAAAACTTGTGAAAAAA
Frob3G15086       -1725 TATTGAGACTATATTTTTGTGGTGGAAATCATATGAATTTATGAAAAACTCATGAAGAAA
Fson3G09193       -3515 TATTGAGACTATATTTTTGTGGTTGAAATGATATGATTTTATGAAAAACTCGTGAAAAAA
                        ***************  * **** ** ** ** *** ******

Ftri3G16655 (1.1)       TTAAATTGGACAGAGGAAATCAAAAACAAAATTGGATCTTTCATAT-CACGAAAAGGCAG
Ftri3G07887 (1.2)       ------------------------------------------------------------
Ftri3G30452 (1.3)       ------------------------------------------------------------
FlinNA39315             TTGAATTAGAAAGAGGAAATAGAAAGCAAAGTTGGATCTTTCATATCCACGAAAAGACAT
Fram3G09579             TTAAATTGGAAAGAGGAAATAGAAAGCAAAATTGGATCTTTCATAT-CACGAAAAGGCAT
Frob3G15086             TTGAATTAGAAAGAGGAAATAGAAAGCAAAGTTGGATCTTTCATAT-CACGAAAAGGCAT
Fson3G09193             TTGAATTGGAAAGAGGAAATAGAAAGCAAAGTTGGATCTTTCATAT-CACGAAAAGGCAT

Ftri3G16655 (1.1)       TAGTTCTTGCCACTTGACCAAGGAGTGTTCGTAGAGCCGTACTTACTCACTAAAACAAAC
Ftri3G07887 (1.2)       --------------------------------AGAGCCATACTTAGTCTCTAAAACAAAC
Ftri3G30452 (1.3)       --------------------------------AGAGCTGTACTTACTCACTAAAACAAAC
FlinNA39315             GA--TTTTGCCACTTGACCAAGGAGTGTTCGTAGAGCCGTACTGACTCACTAAAACAAAC
Fram3G09579             GAATTCTTGCCACTTGACCAAGGAGTGTTCGTAGAGCCGTACTTACTCACTAAAACAAAC
Frob3G15086             GAGTTCTTGCCACTTGACCAAGGAGTGTTCGTAGAGCCGT-CTTACTCACTAAAACAAAC
Fson3G09193             GAGTTCTTGCCACTTGACCAAGGAGTGTTCGTAGAGCCGTACTTACTCACTAAAACAAAC
                                                        *****  * ** * ** ***********
```

MEM1 A submodule
note: Ftri3G16655: ancestral copy; Ftri3G30452: reported copy.
MEM1 B submodule

Fig. S 13. **The Mesophyll expression module 1 in the three Ftri*PEPC1*s**

The mesophyll expression module 1 (MEM1) sequences of the *PEPC1* promoters from five *Flaveria* species. The A and B submodules are highlighted in red boxes. Asterisks show identical nucleotides across the two modules. Pink zones indicate single nucleotide difference in the A submodule, and the required tetranucleotide CACT in the B submodule. In Ftri, *Ftri3G1665* (*PEPC1.1*) is the ancestral copy and the other two are latterly duplicated. The duplicated *PEPC1*s show a segment of deletion between A submodule and B submodule. (Abbreviations: MEM: mesophyll expression module, which is required for the mesophyll highly expression of *PEPC* in *Flaveria* C_4 species as reported in a previous study [22])

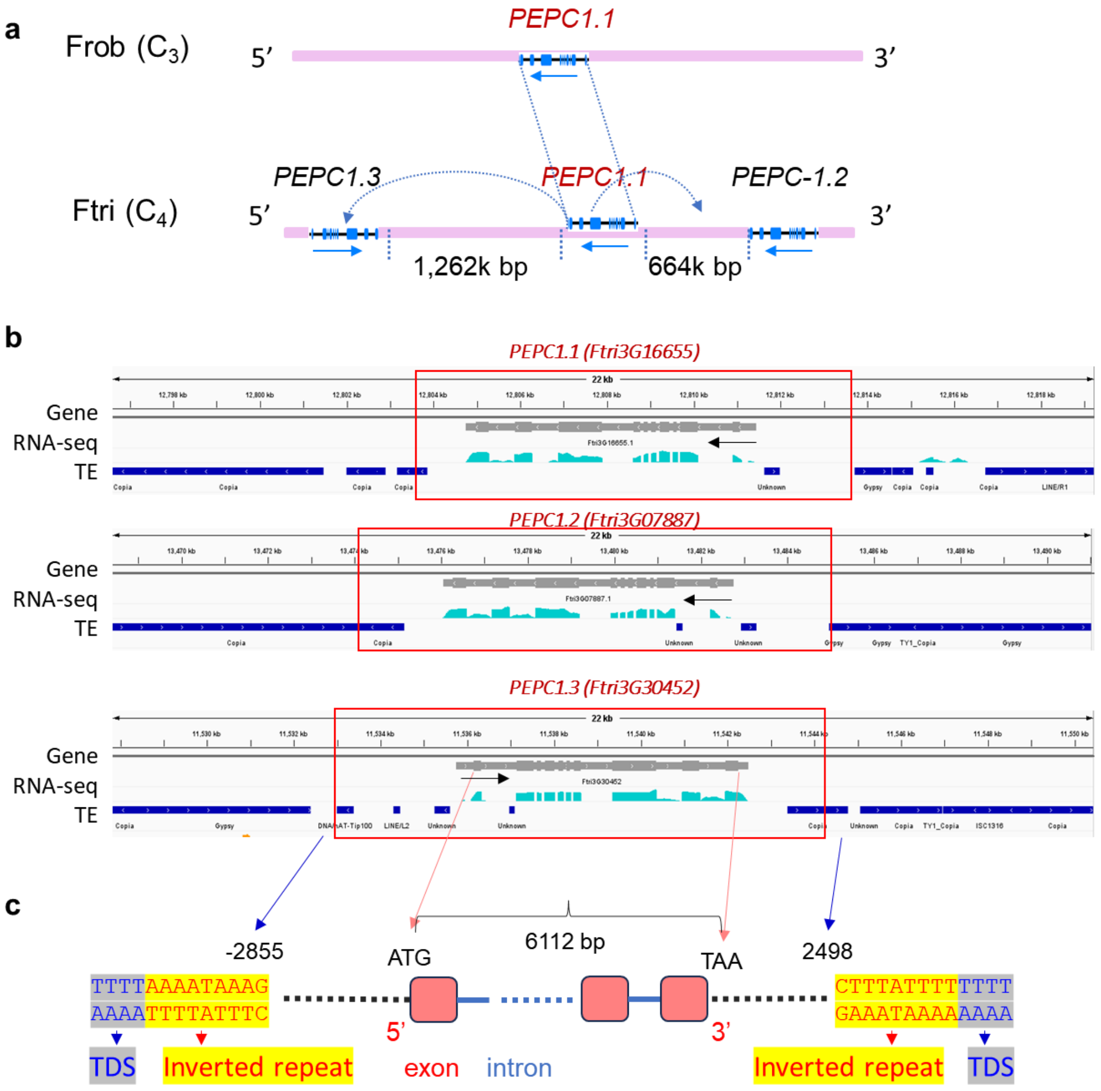


Fig. S 14**. Tandem duplicated Ftri*PEPC1*s may be mediated by retrotransposons**

(a) Schematic illustration of the transposition of Ftri*PEPC1*. (b) Integrative genomics viewer (IGV) indicating genes, transcript abundances and predicted TEs on the chromosomal regions near three Ftri*PEPC1*s. The conserved regions of the three Ftri*PEPC1s* were shown in red frame. (c) Inverted repeats (red font in yellow background) were observed adjacent to the conserved region of Ftri*PEPC1.3*. A 4-bp motif (blue font in grey background) flanks the inverted repeats, resembling a target site duplication (TSD) in transposition event mediated by retrotransposons.

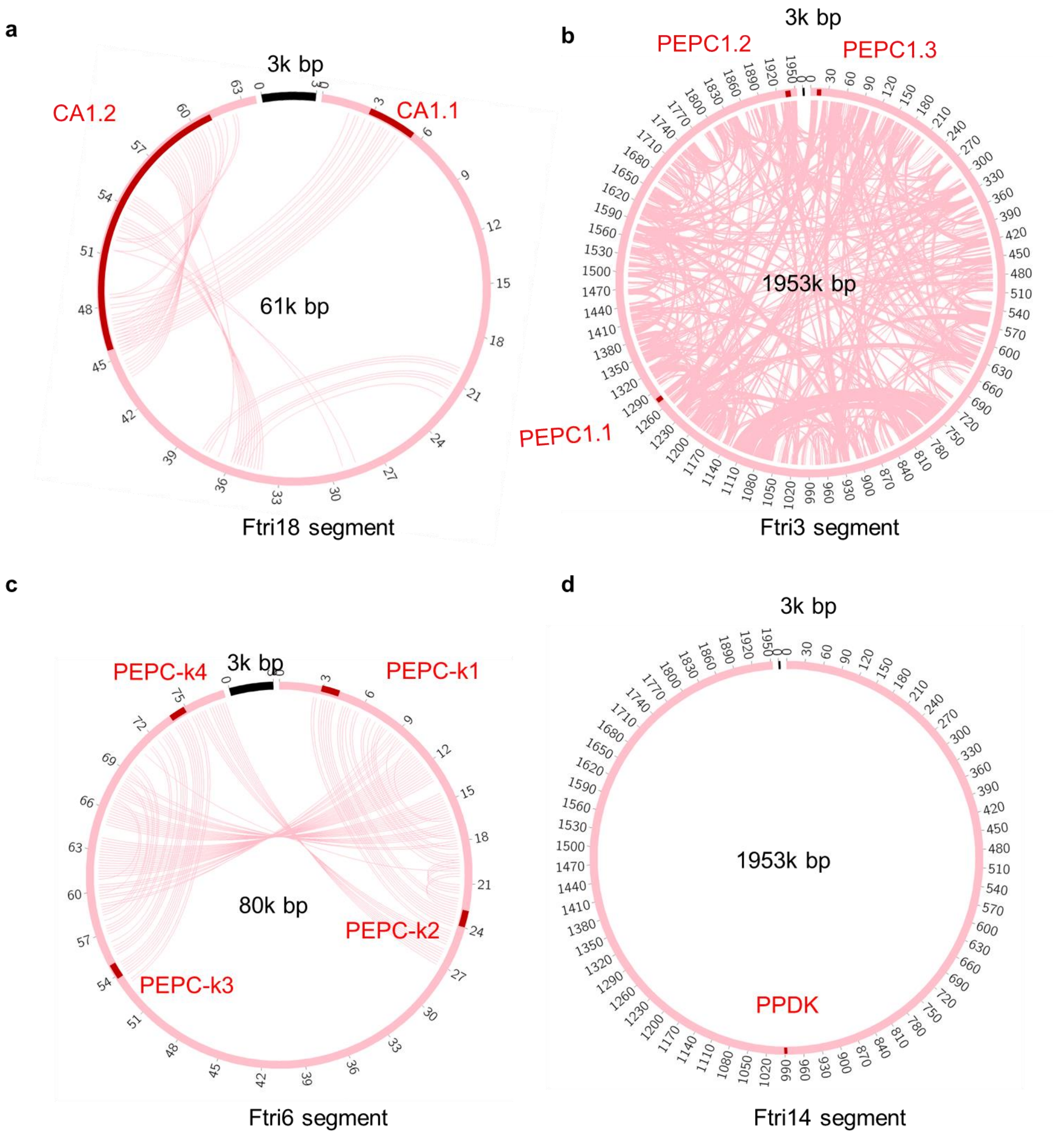


Fig. S 15**. Chromosome segments containing tandem duplicated C_4 gene copies in Ftri**

The chromosome segment containing tandem duplicated *CA1*(a), *PEPC1* (b) and *PEPC-k1* (c) in Ftri. The lengths of shown chromosome segments were as indicated in the circles. (d) The chromosome segment containing *PPDK* in Ftri, which showed a length of 1953k bp of chromosome 14. We used a length of 1953kbp to compare against chromosomal segment on chromosome 13 containing three *PEPC1*s in (b). The 3kbp region in black is used as a scale of the chromosome segment size. Lines inside the circles represent duplicated segments with 200bp. Note that abundant duplicated sequences were observed in the chromosome regions containing duplicated paralogs of *CA1*, *PEPC1* and *PEPC-k1*, but not in the chromosome region containing *PPDK*, which is a singleton gene.

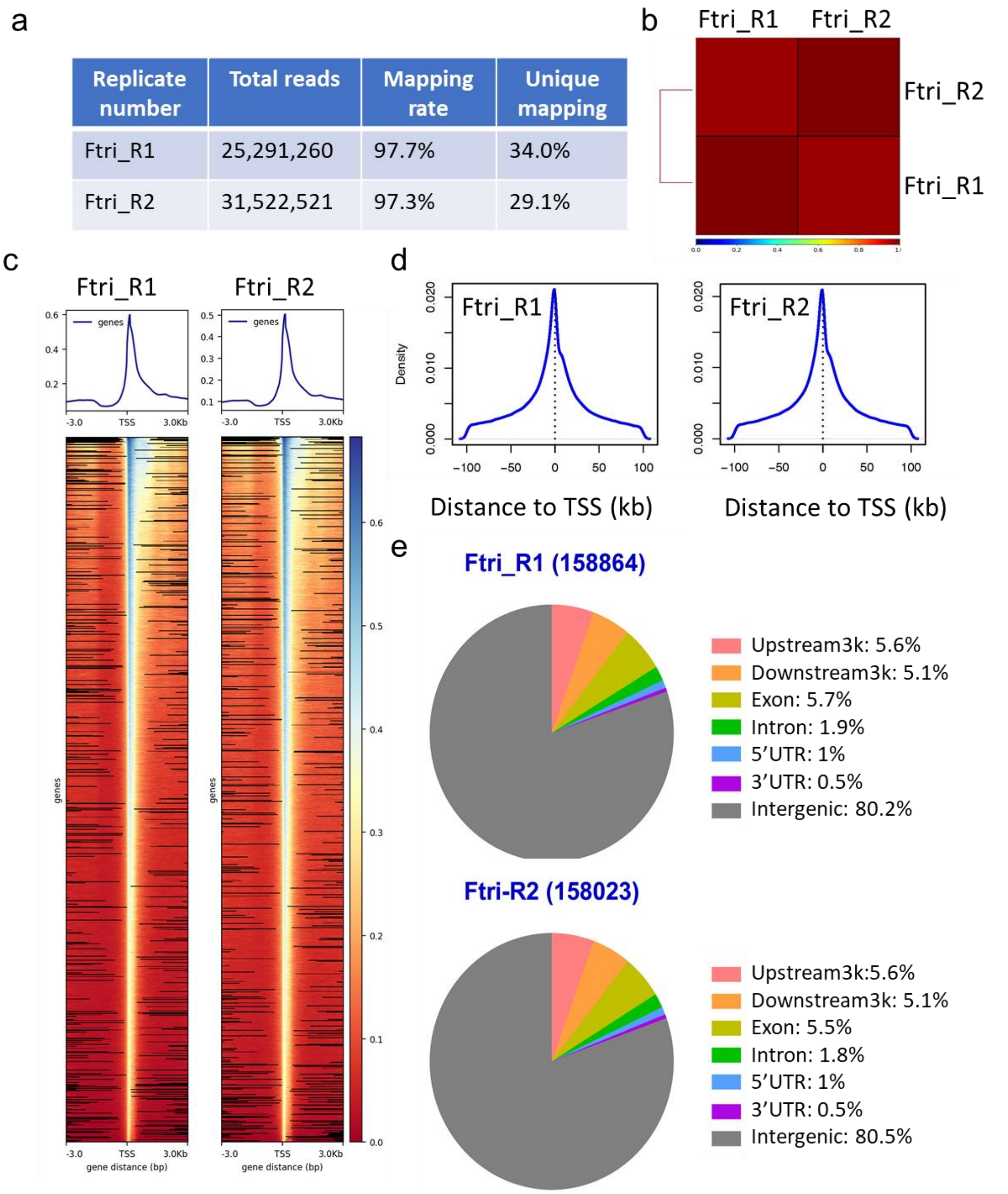
10. Prediction of *cis*-regulatory elements using ATAC-seq

Methods

Nucleic isolation, library construction sequencing, peaks calling and Tn5 hypersensitive site (THS) were illustrated in the Methods of the manuscript.

Results

Approximately 30% ATAC-seq raw reads were uniquely mapped to the genome sequences of Ftri. A high Pearson correlation was reported between two independent biological replicates of ATAC-seq mapping. Around 80% of mapped reads were distributed in the intergenic region, and comparable proportions of mapped reads, *i.e.*, ~5%, were distributed in the 3k bps upstream, 3k bps downstream of genes and exon regions (Fig. S 16). Predicted THS of C_4 genes were predicted and illustrated in Fig. S 17.



Replicate number	Total reads	Mapping rate	Unique mapping
Ftri_R1	25,291,260	97.7%	34.0%
Ftri_R2	31,522,521	97.3%	29.1%

Fig. S 16. **Statistics of ATAC-seq mapping and THS calling**

(a) Table showing the total reads and mapping rates of two biological replicates. (b) Pearson correlation of reads mapping profiles of the two biological replicates. (c) Heatmaps showing the distribution of predicted Tn5 hyper sensitive site (THS) from 3k bps upstream and downstream of the transcript start site respectively. (d) The distribution of predicted THS from 100 k bps upstream and downstream of the transcript start site respectively. (e) Pie charts showing the relative locations of THS to genes. (Abbreviations: ATAC-seq: transposase-accessible chromatin using sequencing; THS: Tn5 hyper sensitive site; TSS: transcript start site.)

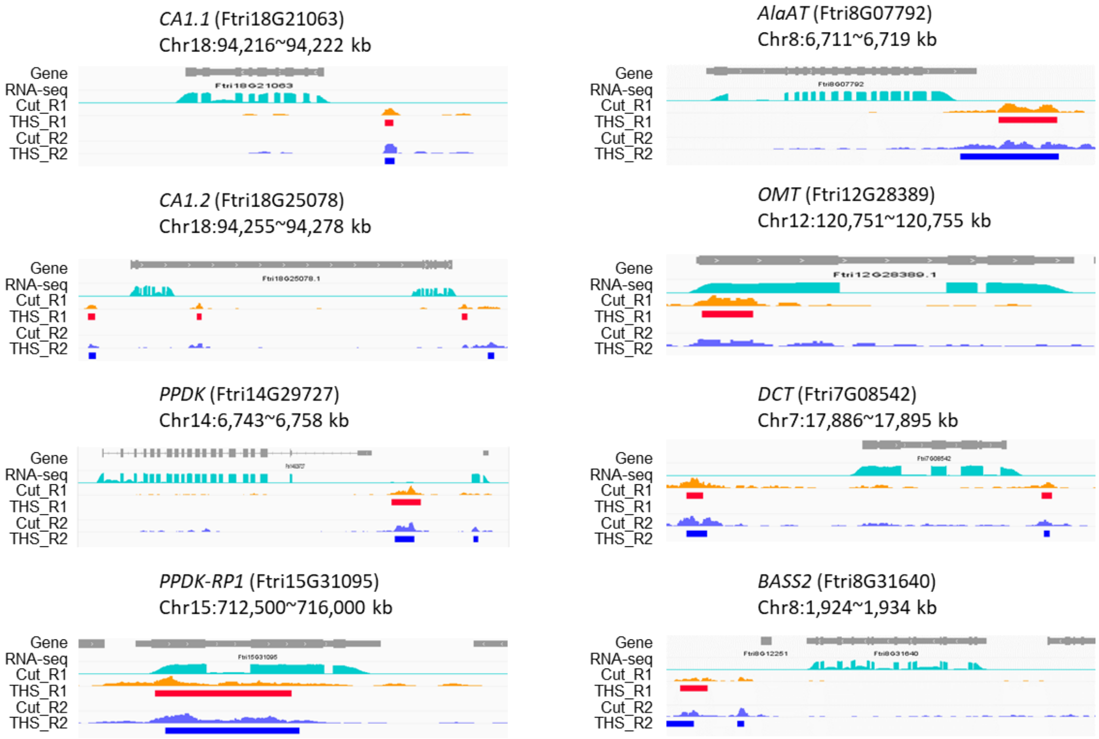


Fig. S 17. **Predicted Tn5 hypersensitive sites associated with C_4 genes in the C_4 species Ftri**

Integrated Genome Viewer (IGV) of RNA-seq reads and ATAC-seq reads of C_4 genes anchored to chromosomes in Ftri. Tn5 cuts and transposase hypersensitive sites (THS) from two biological replicates are shown.

11. ERF *cis*-regulatory elements were abundant in photosynthesis related genes of different C_3 species

Methods

ATAC-seq data of *Zea mays* mesophyll cell and bundle sheath cells were obtained from a previous study [23]. The data of TF binding sites (TFBS) of *Zea mays*, *Setaria italica* and *Sorghum bicolor* were from a prior study [24]. To investigate the enriched CREs of C_4 genes in the C_4 species Ftri, the promoter region (3k bps upstream of the start codon) of C_4 genes were compared with the rest of the genes using findMotifsGenome.pl within Hommer [25] (v4.11.1) with default parameters.

Results

ERF CREs were also the most abundant CREs in *Zea mays* (corn), *Setaria italica* and *Sorghum bicolor* (Fig. S 18 and Fig. S 19).

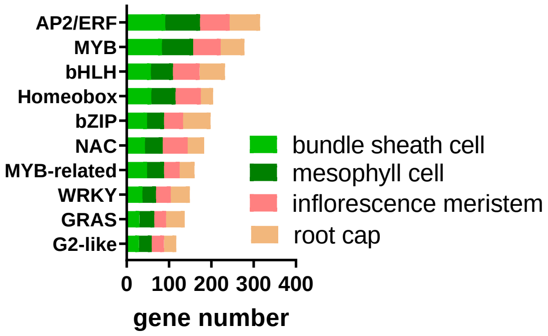


Fig. S 18**. A large number of ERF TFs show positive chromatin accessibility in different tissues of *Zea mays***

Bar plots showing the profiles of chromatin accessibility of TFs from different types of tissues in *Zea mays*. The top ten TFs with the greatest number of genes showing chromatin accessibility are shown here. A large number of ERF TFs show positive chromatin accessibility in different tissues, especially in the mesophyll and bundle sheath. TF frequency data used for drawing this bar plot are from a previous study [23] which are based on single cell ATAC-seq. (Abbreviations: MC: mesophyll cell; BSC: bundle sheath cell.)

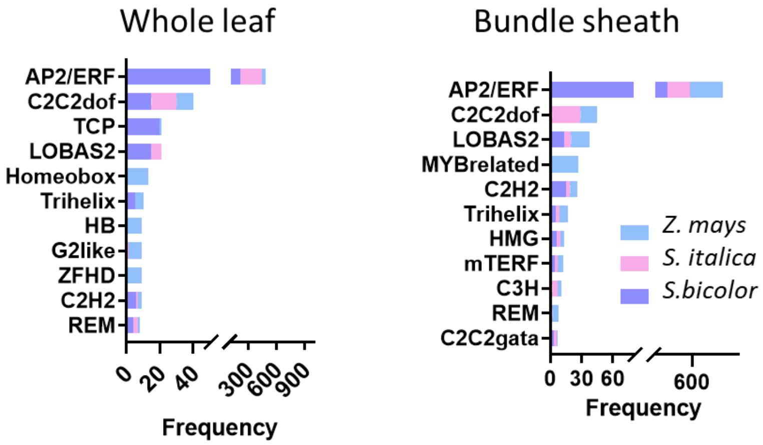


Fig. S 19**. ERF TF binding sites are abundant in photosynthesis related genes**

Bar plots showing the frequency of different TF binding site (TFBS) in photosynthesis related genes in the chromosome regions of the three C_4 species, *i.e.*, *Zea mays* (*Z. mays*), *Setaria italica* (*S. italica*) and *Sorghum bicolor* (*S. bicolor*). Photosynthesis related genes include genes in C_4 pathway, photorespiratory pathway and Calvin-Benson-Bassham cycle. The top 10 most frequently known TFBS are showed. The data of TFBS used for drawing bar plots here are from a previous study [24], in which the chromosome open regions were computed based on DNase-seq.

12. Construction of gene regulatory networks

Methods

We constructed genome-wide co-regulatory networks (GRNs) for each of the five *Flaveria* species, which based on the gene expression profiles of at least 18 RNA-seq datasets either from a previous work [26] (for Frob, Fson, Fram and Ftri) or generated in the current study (for Flin) (Table S 8). The GRNs were constructed following a previously decribed method [26] . Specifically, genome-wide gene regulatory networks (GRNs) of five *Flaveria* species were constructed independently through application of the CMIP software package [27]. The CMIP package implements a path consistency algorithm based on the conditional mutual information which is specifically designed to predict direct regulation between genes [28]. Based on the gene expression table (the only input), the method generates a zero-order network by computing mutual information (MI), and indirect relationships were removed considering the CMI, which resulted in a first-order network. We implemented a *P*-value determination in the CMIP package based on a permutation test. Specifically, we shuffled the expression level of each gene 2000 times, generating 2,000 null datasets. Thereafter, the *P*-value was defined as the ratio of number of null datasets to 2000 whose CMI values were higher than that calculated from the original dataset. Here, a *P*-value of 0.001 was used as the cutoff. The C_4GRN was obtained by retaining the GRN of the C_4 genes and their co-regulated TFs (thereafter, C_4TFs) from the genome-wide GRN.

In the original CMIP package, the cutoff of gene-gene partial Pearson correlation coefficient (PCC) was estimated as the inflection point of the exponential-function-fitted curve with the relationship between edge number and PCC values [27]. Considering that we mainly focused on the C_4GRN, we revised this computational procedure to determine the proper PCC. Specifically, we compared the relative abundance of all TF families (TF frequency) of the C_4GRN obtained using different PCC cutoffs, increasing from 0.1 to 0.9 with a step size of 0.1. We found that the TF frequency maintained a high degree of similarity under different PCC cutoffs ranging from 0.1 to 0.7 (Fig. S 20). We found that a PCC of 0.5 represented the inflection point in the curve of relationship between edge number of C_4GRNs and PCC cutoffs (Fig. S 21). We thus selected 0.5 to be the PCC cutoff for GRN construction.

Results

We constructed a genome-wide co-regulatory network (GRN) of the five *Flaveria* species based on the gene expression profiles of at least 18 RNA-seq datasets either from a previous work [26] or generated in the current study (Table S 8).

Previously, we constructed C_4GRN for Frob, Fson, Fram and Ftri using the same RNA-seq datasets, whereas genes were annotated based on the transcript of Fram [26] (transcriptome-based C_4GRN). We compared the C_4GRN constructed here (genome-based C_4GRN) with transcriptome-based C_4GRN for the four species. The total gene number and TFs were enhanced due to the improved gene annotations, which resulted in increased C_4TFs across all the species (Fig. S 23 a), nevertheless , the trends of the number of C_4TFs along evolution (from C_3 to C_4) were consistent between the two studies[26] . The updated C_4GRN demonstrated that 1/3 C_4TFs were species specific, which was lower than the transcriptome-based C_4GRN (Fig. S 23 b and c), but still consistent with previous report [26] that a large number of TFs were species specific. In addition, among the top 20 TF families with the greatest number of C_4TFs, 15 of them were common across the two studies (Fig. S 23 d and e). On the other hand, the genome-based GRNs were improved as following (1) with the genome annotation, we improved the annotation of TFs for the five Flaveria species sequenced here, (2) combined with *cis*-elements information, we filtered out C_4 genes co-regulated TFs with no predicted cognate *cis*-regulatory elements on the promoter. These improvements enabled us to find out that ERF TFs were more abundant in C_4GRNs in the C_4 species than in other species (Fig. S5 in the manuscript). Moreover, the gene structure annotation enabled us to uncover that intronless ERF TFs were abundant in C_4GRNs of the C_4 species, which can not be figured out based on the previously general GRNs (Fig. S5 in the manuscript).

While the total number of ERF were comparable in all the five *Flaveria* species, C_4GRN of the C_4 species Ftri showed more ERF TFs than those in other *Flaveria* species (Fig. S 22), specifically, 27 ERF TFs were predicted in Ftri C_4GRN (Fig. S6 c in the manuscript) compared to 11 in Frob C_4GRN (Fig. S 24 a). When only one copy of *CA1*, *PEPC1* and *PEPC-k* were analyzed for C_4GRN in the C_4 species Ftri, 22 ERF TFs were co-regulated with 13 C_4 genes (Fig. S 24 b).

We examined how many of the counterparts of the 27 C_4 ERF TFs in the C_3 species Ftri also regulated the counterparts of C_4 genes in the C_3 species Frob, and the results showed the number was four (Table S 9), suggesting that most C_4 ERF TFs were recruited to regulate C_4 genes in the later stage of C_4 evolution, which is consistent with our earlier study [29].

Next, we examined how many of the shared TF between Ftri and Zmay co-regulate C_4 orthologous genes in dicot C_3 species Frob (from this study) and monocotyledonous C_3 species *Oryza sativa* (Osat). The GRN of Osat was constructed based on RNA-seq data using the same package (PCA-CMI) [27] with Spearman correlation coefficient >0 .8, Pearson correlation coefficient > 0.8 and P-value < 0.001(The Osat of GRN was not published yet). The RNA-seq data from Osat was obtained from 16 time points within one day, with three replicates for each time point. We found that among shared ERF TFs between Ftri C_4GRN and Zmay C_4GRN (including 14 Ftri ERF TFs and 12 Zmay TFs), 3 of them were in the C_4GRN of C_3 species Frob, and 2 of them co-regulated C_4 orthologous genes in Osat (Table S 10). The results therefore suggested that common TFs between Ftri and Zmay were recruited to regulate C_4 genes during the later stage of C_4 evolution, i.e., they do not regulate C_3 counterparts in C_3 ancestors.

Though the ERF TFs were present in C_3 species, their expression level changed during C_4 evolution. Among the 323 TFs in the C_4GRN of the Ftri (C_4), 96 TFs (including 7 ERF TFs) and 55 TFs (including 1 ERF TF) showed significantly higher and lower transcript abundance in Ftri than in Frob (C_3), respectively (Fig. S 25 a). Besides, among the shared 14 ERF TFs between Ftri C_4GRN and Zmay C_4GRN, five of them showed higher transcript abundances in Ftri (C_4) than their counterparts in Frob (C_3) (Fig. S 25 b), suggesting that though C_4 ERF TFs were present in C_3 species, the expression level of ERF TFs were changed during evolution.

Table S 8**. RNA-seq datasets used for constructing gene regulatory networks for the five *Flaveria* species**

Species	**low CO_2**	**low CO_2**	**low CO_2**	**ABA**	**high light**	**# total Samples**
	2 weeks	4 weeks	6 months	3 hours	2 weeks	
Frob	6	6	4	6	6	**28**
Fson	6	6	4	6	n.a	**22**
Flin	6	6	n.a	n.a	6	**18**
Fram	6	6	n.a	6	6	**24**
Ftri	6	6	n.a	6	6	**24**

Note: n.a: not available

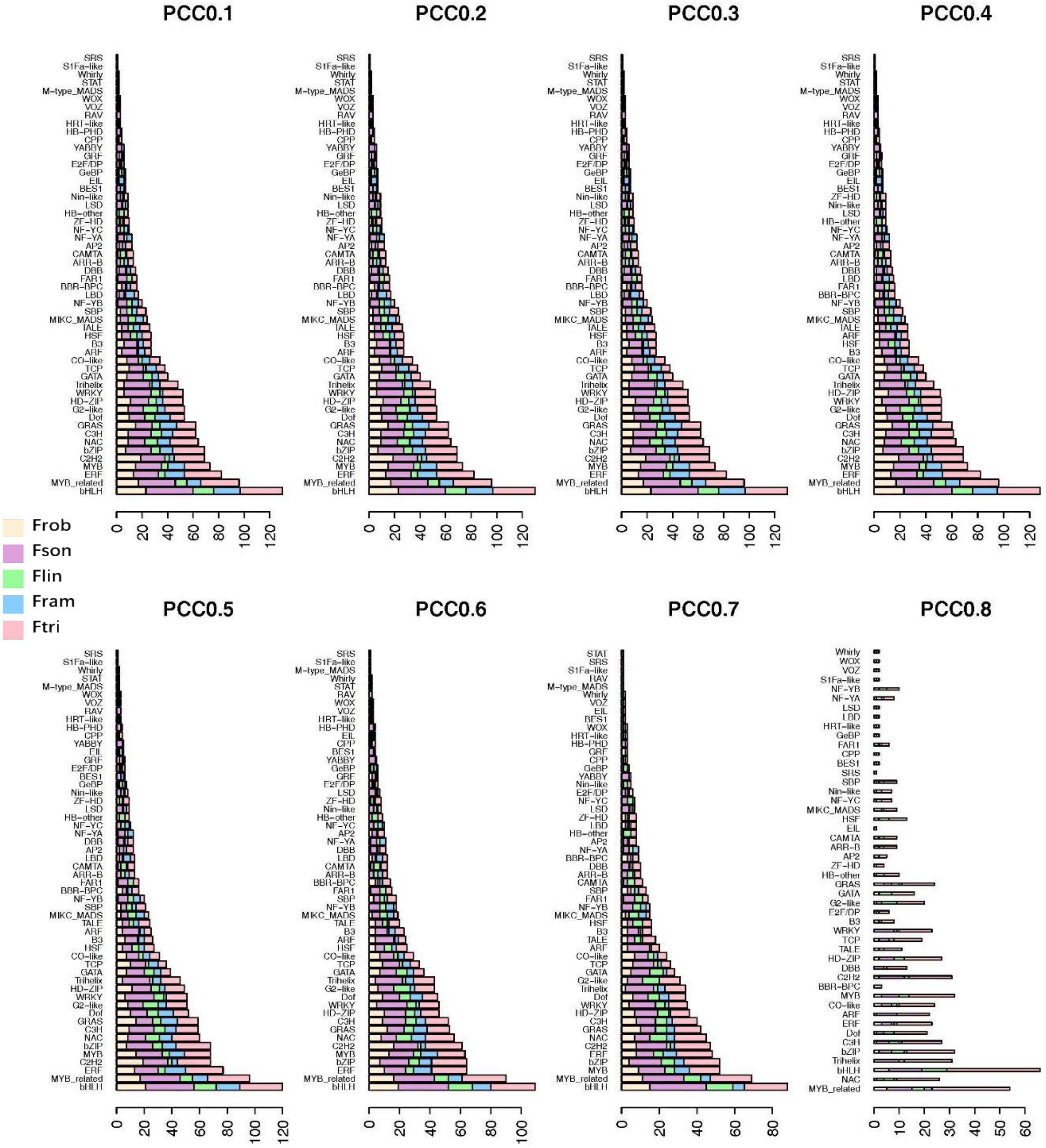


Fig. S 20**. Distributions of** C$_4$**TFs in each TF family for each** ***Flaveria*** **species under different PCCs cutoffs**

Bar plots showing the TFs number for each TF family across four *Flaveria* species under different thresholds of PCC cutoffs ranging from 0.1 to 0.8. The X axes display the number of TFs for each TF family.

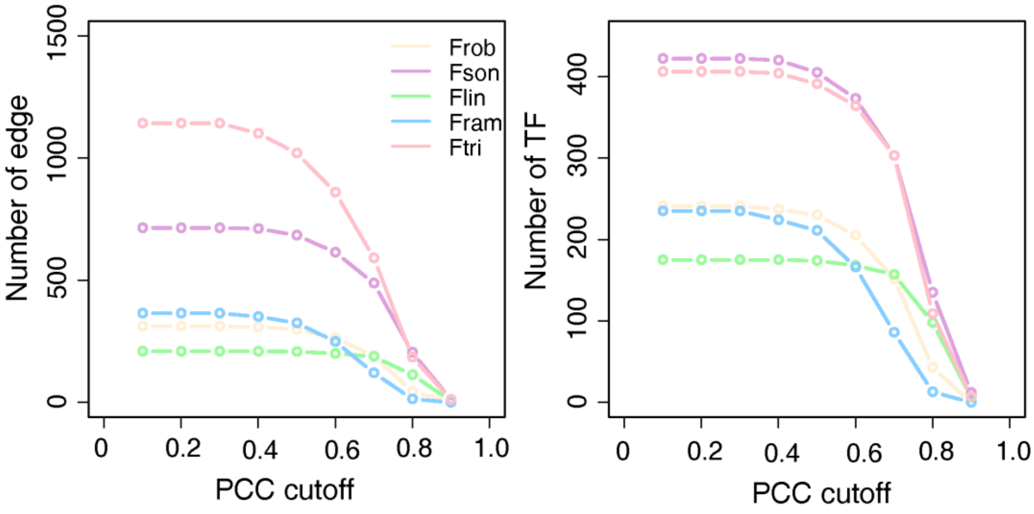


Fig. S 21**. The number of edges and TFs of** C4**GRN under different PCC cutoffs**

Curves showing the distributions of C_4GRN edges (interaction between C_4 genes and TFs) in five *Flaveria* species under PCC cutoffs ranging from 0.1 to 0.9. The number of edges shows an inflection point at a PCC of 0.5.

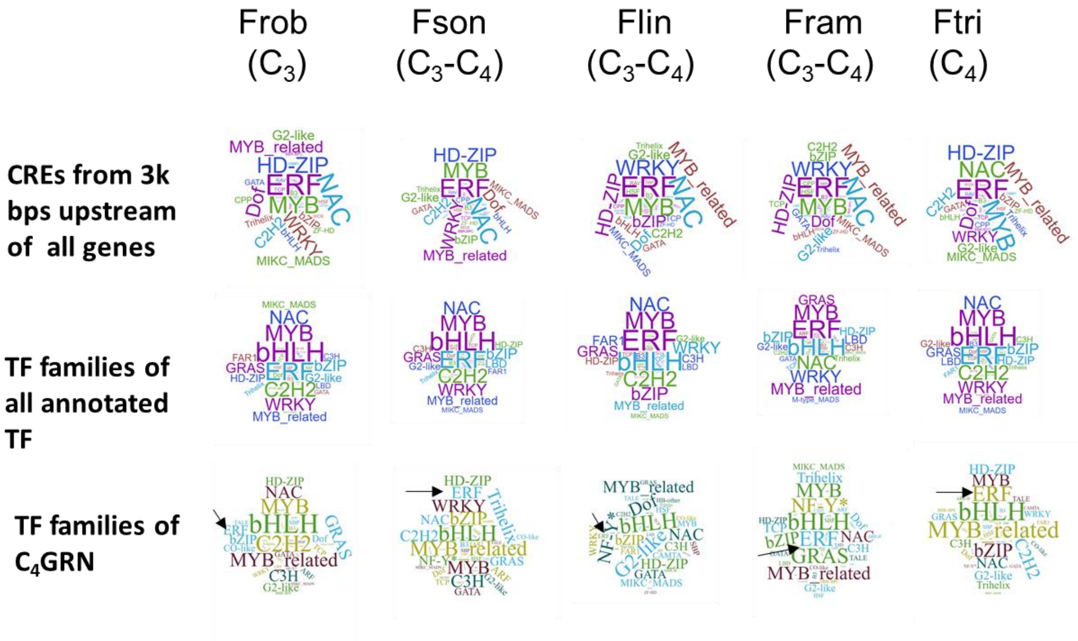
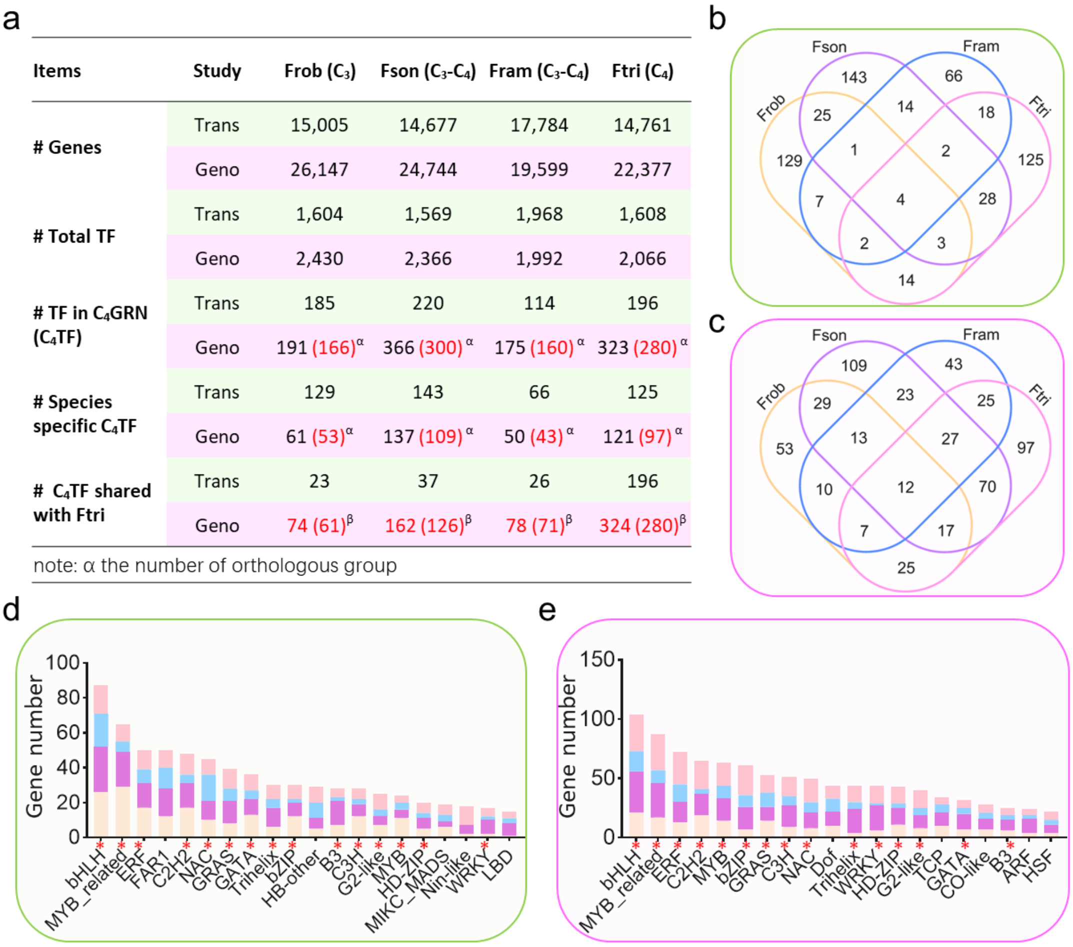


Fig. S 22**. ERF TFs are more abundant in** C4**GRN of the** C4 **species than those of other species**

Word clouds indicating the frequency of predicted *cis*-regulatory elements (CREs) from the 3k bps upstream of all genes in different species (top panel); all predicted TFs (middle panel) and C_4 genes co-regulated TFs (bottom panel), with larger names indicating more abundance. Note that ERF CREs are abundant across all the five *Flaveria* species, and the number of annotated ERF TFs are comparable in the five *Flaveria* species, whereas ERF TFs are more abundant in C_4GRN of the C_4 species Ftri than those of other species.

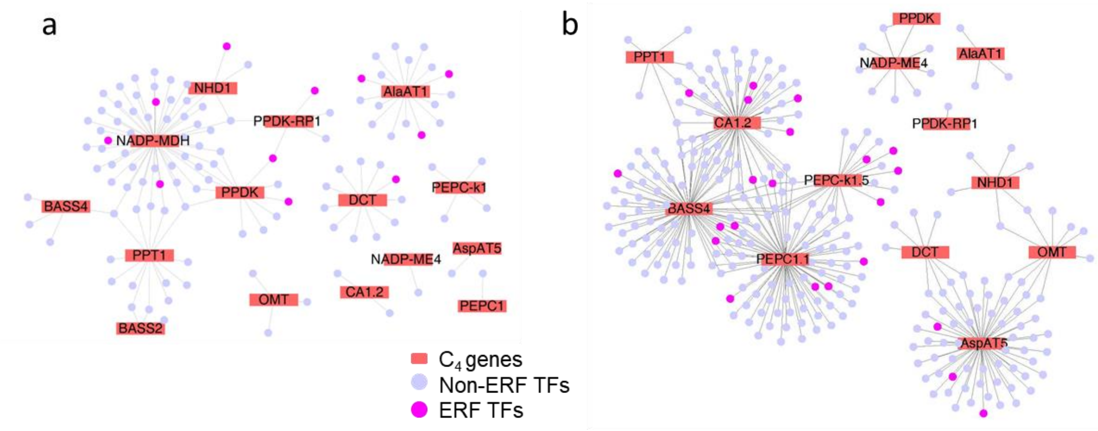


Items	Study	Frob (C$_3$)	Fson (C$_3$-C$_4$)	Fram (C$_3$-C$_4$)	Ftri (C$_4$)
# Genes	Trans	15,005	14,677	17,784	14,761
	Geno	26,147	24,744	19,599	22,377
# Total TF	Trans	1,604	1,569	1,968	1,608
	Geno	2,430	2,366	1,992	2,066
# TF in C$_4$GRN (C$_4$TF)	Trans	185	220	114	196
	Geno	191 (166)$^{\alpha}$	366 (300)$^{\alpha}$	175 (160)$^{\alpha}$	323 (280)$^{\alpha}$
# Species specific C$_4$TF	Trans	129	143	66	125
	Geno	61 (53)$^{\alpha}$	137 (109)$^{\alpha}$	50 (43)$^{\alpha}$	121 (97)$^{\alpha}$
# C$_4$TF shared with Ftri	Trans	23	37	26	196
	Geno	74 (61)$^{\beta}$	162 (126)$^{\beta}$	78 (71)$^{\beta}$	324 (280)$^{\beta}$

note: α the number of orthologous group

Fig. S 23. Comparison of C$_4$ GRN based on RNA-seq data with either genome annotation or transcript annotation.

(a) Statistics of GRNs based on transcript-based annotation (Trans. for short) from a previous study [26] and genome-based annotation (Geno. for short) from this study. α and β represent the number of orthologous groups that the under-studied TFs belonging to. (b) and (c) Venn diagram showing the intersection between TFs across four *Flaveria* species from transcript-based C$_4$GRN and genome-based C$_4$GRN, respectively. The numbers in (c) represents the number of orthologous groups instead of genes. (d) and (e) Bar plots showing the frequency of the top 20 TF families with the greatest number of genes in C$_4$GRNs from transcript-based GRN and genome-based GRN respectively. TF families shared between transcript-based GRN and genome-based GRN are marked with red a star.



Frob (C_3): 11ERF TFs /15 C_4 genes

Ftri (C_4): 22 ERF TFs/13 C_4 genes

Fig. S 24. The C_4GRN of Ftri and Frob

(a)The C_4GRN of the C_3 species Frob. The C_4GRN of the Ftri (C_4) was shown in Fig. S6 in the manuscript. (c) C_4GRN of Ftri including only one copy for *CA1*, *PEPC1*, and *PEPC-k1* respectively.

Table S 9. **Four counterparts of C_4 ERF TF in Ftri coregulate counterparts of C_4 genes in the C_3 species Frob**

Orthologous group	ID of Frob (C_3)	TF family	Coregulated C_4 orthologous genes in C_4GRN
OG0003450	Frob15G32959	ERF	Frob9G16612_NADP-MDH
OG0003676	Frob11G13246	ERF	Frob9G30036_NADP-ME4
OG0001395	Frob18G09268	ERF	Frob3G31625_NHD1
OG0001395	FrobNA15531	ERF	Frob9G30036_NADP-ME4

Table S 10**. Common C_4 ERF TF between Ftri and Zmay that are coregulated with counterparts of C_4 genes in the C_3 species Frob and Osat**

Orthologous group	ID in C_3 species	TF family	Coregulated C_4 orthologous genes
Frob (C_3)			
OG0003450	Frob15G32959	ERF	Frob9G16612_NADP-MDH
OG0001395	Frob18G09268	ERF	Frob3G31625_NHD1
OG0001395	FrobNA15531	ERF	Frob9G30036_NADP-ME4
Osat (C_3)			
OG0000014	Os01g0752500	ERF	Os01g0752500_PPDK-RP
OG0000023	Os01g0797600	ERF	Os01g0797600_DIC

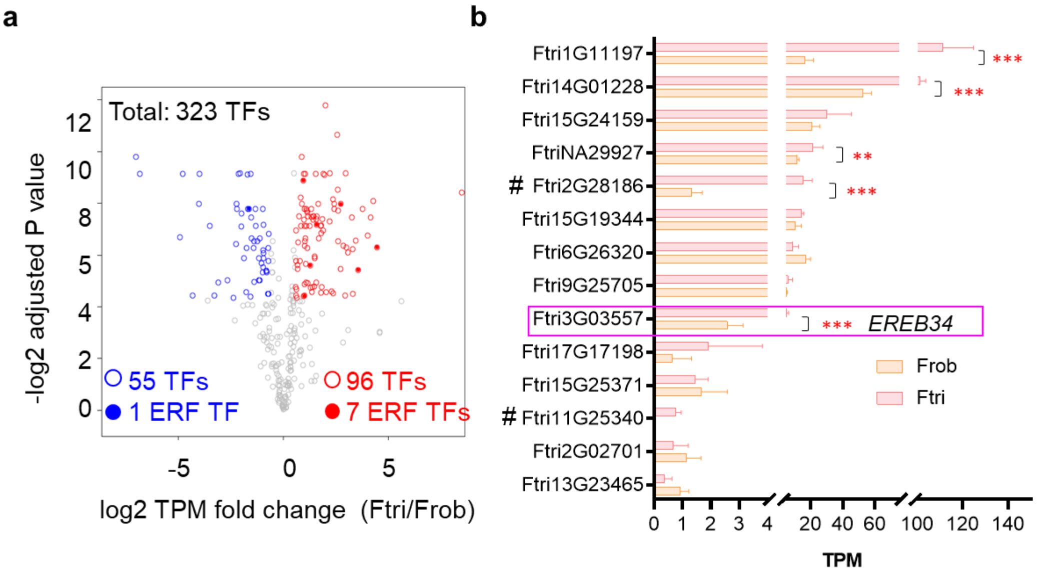


Fig. S 25. The change of transcript abundances of C_4TF between Ftri and Frob

(a) Volcano plot shows differentially expressed C_4TFs of Ftri (C_4) between Ftri and Frob (C_3). (b) The transcript abundance of 14 common C_4 ERF TFs between Ftri and Zmay in Ftri and their counterparts in Frob. Intron-containing genes are marked with "#", *EREB34* is indicted. Statistical significance between Frob and Ftri were calculated using edge R with an adjusted *P* value threshold of 0.05 and a foldchange threshold of 1.5. Statistical significances were represented as *: *P*<0.05, ** *P*<0.01, and *** *P*<0.001. (Abbreviation: TPM: transcript per million mapped reads.)

13. Comparison of transcript abundances based on RNA-seq data

<u>Methods</u>

RNA-seq data from Frob, Fson, Fram and Ftri are from our previous study [30] (Table S 8). Additionaly, 18 RNA-seq datasets of Flin were generated in this study, including six RNA-seq datasets from a 2-week low CO_2 experiment (three from normal CO_2 condition, *i.e.*, 380 ppm, and three from low CO_2 condition *i.e.*, 100 ppm), six RNA-seq samples from a 4-week low CO_2 experiment (three from normal CO_2 condition and three from low CO_2 condition), and six RNA-seq datasets from high light experiment (three from normal light condition, *i.e.*,500 µmol m^{-2} s^{-1}, and three from high light condition, *i.e.*, 1400 µmol m^{-2} s^{-1}). The low CO_2 experiments on Flin were conducted with those of the other four species. Plants were grown in a growth chamber with a photosynthetic photon flux density (PPFD) controlled to be 200 µmol

m^{-2} s^{-1}, and a temperature of 22 ±2 °C, 70% relative humidity (RH), and a photoperiod of 16 hours light/8 hours dark. The CO_2 concentration used for the low CO_2 treatment was 100 ppm and the control CO_2 concentration was 380 ppm. The high light experiment on Flin was conducted together with those of Frob, Fram and Ftri. Plants were grown in a phytotron with a PPFD of 1400 µmol m^{-2} s^{-1} for high light condition and 500 µmol $m^{-2}s^{-1}$ for normal light condition. The high light condition was achieved by supplementing light with a lab-made light emitting diode (LED) light source. Genes were clustered into 10 patterns using Kmeans based on their transcript abundance across the five *Flaveria* species.

Results

To compare transcript profiling of the five *Flaveria* species, three RNA-seq datasets from the normal CO_2 condition of the 2-week low CO_2 experiment and three from the normal CO_2 condition of the 4-week low CO_2 experiment were used for all five *Flaveria* species. Transcript abundances were calculated as transcript per million mapped reads (TPM). In total, 27,684 genes were retained with the maximum transcript abundance no less than 1 TPM in the five *Flaveria* species. We found that six replicates of a certain species had high Pearson correlations (Fig. S 26). The 27,684 genes were clustered into 10 patterns according to their evolutionary profiles (Fig. S 27). C_4 genes were mainly clustered in one pattern (P3), in which C_4 species showed the highest transcript abundances, and photorespiratory genes were mainly clustered in one pattern (P8), where Flin showed the highest transcript abundance.

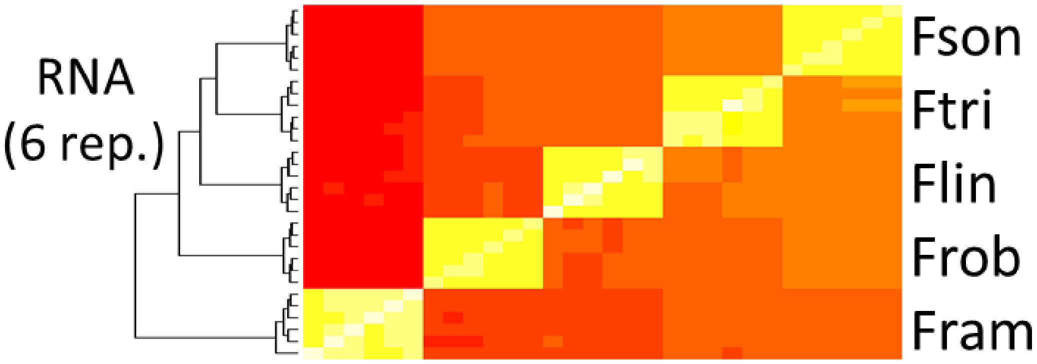


Fig. S 26. **Pearson correlations of RNA-seq datasets for five *Flaveria* species in transcript abundances**

(a) Heatmap showing the Pearson correlations of RNA-seq samples based on transcript abundances of all detected genes. The transcript abundances were calculated as transcript per million mapped reads (TPM). Six RNA-seq samples were used to compare the gene expression profiles along evolutionary history, including three RNA-seq samples from normal CO_2 condition of 2-week low CO_2 experiment, and three RNA-seq samples from normal CO_2 condition of 4-week low CO_2 experiment.

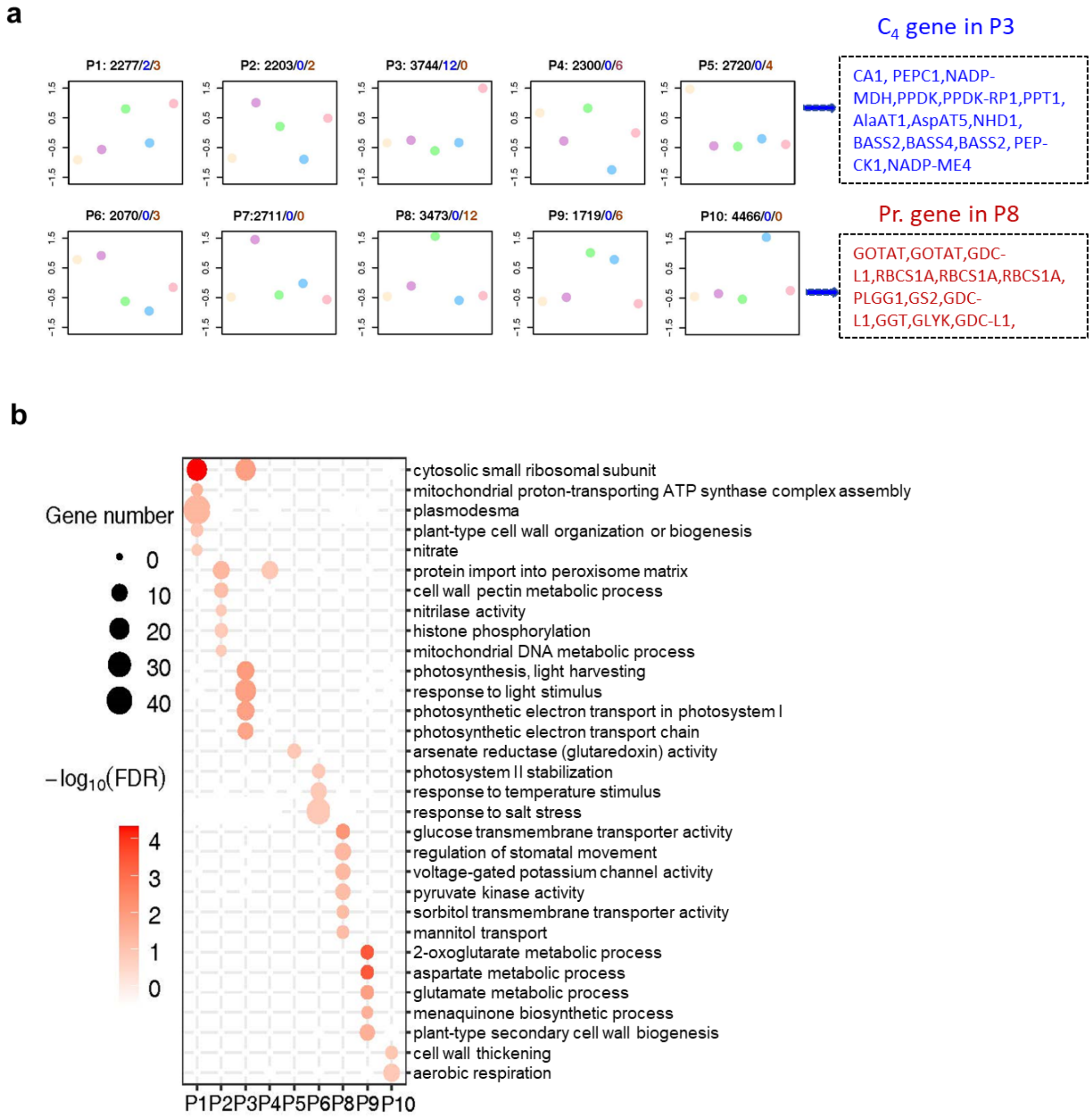


Fig. S 27. **Ten evolutionary patterns of the detected genes based on transcript abundances** (a) Genes were clustered into ten patterns (P for short, from P1 to P10) according to their expression profiles in the five *Flaveria* species. Genes were clustered using Kmeans based on mean transcript abundances from six RNA-seq samples for each species (see methods above). The numbers on the top of each pattern represent: black for all genes, blue for C_4 genes and red for photorespiratory genes. C_4 genes in pattern 3 (P3) and photorespiratory genes in pattern 8 (P8) are listed in dashed frame. (b) The enriched gene ontology (GO) terms of genes from each pattern. Pattern 7 (P7) showed no enriched gene ontology. Enriched GO terms were calculated using a *Fisher's* test with a *P* value threshold of 0.01 adjusted using Benjamini & Hochberg method.

14. Comparison of protein abundance based on proteomics

Methods

Tryptic digestion

A FASP (Filter-Aided Sample Prep) digestion was adapted for the following procedures in Microcon PL-10 filters. After buffer displacement three times with 8 M Urea in 25 mM ABC (ammonium hydroxide), proteins were reduced using 10 mM DTT at 37 °C for 30 minutes, followed by alkylation with 30 mM iodoacetamide at 25 °C for 45 minutes in the dark, and digestion with trypsin (enzyme/protein at a 1:50 ratio) at 37 °C for 12 h after washed with 20% ACN and buffer displaced three times with digestion buffer (30 mM ABC). After digestion, the solution was filtered, and the filter was washed twice with 15% CAN. All filtrates were pooled and vacuum-dried.

High pH HPLC Fractionation

To generate a data dependent acquisition (DDA) library, peptides were prefractionated using a Dionex UltiMate 3000 HPLC system (Thermo Fisher Scientific, USA) with a C18 column (3 μm, 2× 150 mm, Phenomenex, USA). HPLC solvent A was 10 mM NH_4HCO_3, solvent B was 10 mM NH_4HCO_3 in 80% ACN. Peptides from each sample were mixed (a total of 200 μg), dried, and dissolved with 10mM NH_4HCO_3. The mixture was separated by a linear gradient with a flow rate of 200nL/min. The gradient was set as follows: 5-40% in 25 min, and 40-100% in 5 min. Finally, 30 fractions were mixed into 15 samples, and vacumm-dried completely.

LC-MS/MS Analysis

LC-MS/MS analysis was performed using an EASY-nLC 1200 system (Thermo Fisher Scientific, USA) coupled to an Orbitrap Fusion Lumos mass spectrometer (Thermo Fisher Scientific, USA). Samples were resuspended in 1% FA, and iRT peptides (Biognosys, Switzerland) were added prior to MS analysis. Peptides were analyzed using a home-made C18 analytical column (75 µm i.d. × 25 cm, ReproSil-

Pur 120 C18-AQ, 1.9 µm (Dr. Maisch GmbH, Germany)). The mobile phases consisted of Solvent A (0.1% formic acid) and Solvent B (0.1% formic acid in 80%ACN). The peptides were eluted using the following gradient: 2-5% B in 2 minutes, 5-35% B in 100 minutes ,35-44% B in 6 minutes, 44-100% B in 3 minutes, 100% B for 10 minutes, at a flow rate of 200 nL/min.

For DDA experiments, the resolution of full MS scans was set as 60000 at m/z 200, the AGC target was set as 4^{e5} with a maximum injection time of 50 ms. The scan range was set as 400-1200 m/z. For MS2, the Normalized AGC target was set as 100% with a resolution of 15,000 and maximum injection time of 22 ms, and the NCE was set as 30%. The cycle time was set as 2 seconds.

Data independent acquisition (DIA) analysis was set as three full MS scans, and each full MS scan was followed by 20 MS2 windows. The first 20 windows were set as shown in Table S 11; the second 20 windows were set as shown Table S 12; and the third was set as shown Table S 13. The resolution of full MS scans was set as 120000 at m/z 200 and full MS Normalized AGC target was 100% with a maximum injection time of 50 ms, and a scan range was of 400-1200. The resolution of MS2 was set as 30000 with a maximum injection time of 54 ms, and the NCE was set as 32%.

Data Processing

DIA data analysis was performed using Spectronaut (version 14.7, Biognosys, Zurich, Switzerland). High precision iRT calibration was employed. The library was generated by the search engine platform Pulsar in Spectronaut (version 14.7, Biognosys, Zurich, Switzerland) using default settings, and DIA data were analyzed using default settings disabling the PTM localization filter. Mass tolerance/accuracy for precursor and fragment identification was set to default settings. FDR at the peptide and protein level was set to 1% using a mutated decoy model. For matching of DIA data to the spectral library, the applied mass and retention time tolerances (for both MS1 and MS2) are dynamic based on the m/z of the targeted ion and the retention time of the scan. The calibration was performed for each run individually. Default settings for quantification at MS1 level were employed for quantification.

Table S 11**. Settings of the first 20 windows of the MS scan for DIA analysis**

m/z	z	t start (min)	t stop (min)	Isolation Window (m/z)	Normalized AGC Target (%)
403.9562	2	0	120	10.50613	100
412.7018	2	0	120	8.984924	100
420.2895	2	0	120	8.190521	100
427.8015	2	0	120	8.833588	100
435.2217	2	0	120	8.006775	100
442.1428	2	0	120	7.835449	100
449.1415	2	0	120	8.161926	100
455.9776	2	0	120	7.510254	100
462.5025	2	0	120	7.539642	100
469.2666	2	0	120	7.988556	100
476.0137	2	0	120	7.505493	100
482.5152	2	0	120	7.497589	100
489.0068	2	0	120	7.485657	100
495.4973	2	0	120	7.49527	100
502.3288	2	0	120	8.167694	100
509.1054	2	0	120	7.38559	100
515.7923	2	0	120	7.988098	100
522.9423	2	0	120	8.311951	100
529.9496	2	0	120	7.702576	100
537.0469	2	0	120	8.492188	100

Table S 12**. Settings of the second 20 windows of the MS scan for DIA analysis**

m/z	z	t start (min)	t stop (min)	Isolation Window (m/z)	Normalized AGC Target (%)
544.5246	2	0	120	8.463196	100
551.7519	2	0	120	7.991333	100
558.9966	2	0	120	8.497986	100
566.2987	2	0	120	8.106323	100
573.6986	2	0	120	8.693542	100
581.1627	2	0	120	8.234558	100
588.5497	2	0	120	8.53949	100
596.3158	2	0	120	8.992615	100
604.7009	2	0	120	9.77771	100
613.4788	2	0	120	9.777954	100
622.3413	2	0	120	9.947144	100
631.3462	2	0	120	10.06268	100
640.3571	2	0	120	9.959106	100

649.5827	2	0	120	10.492	100
658.9945	2	0	120	10.33167	100
668.4948	2	0	120	10.66895	100
678.3311	2	0	120	11.00366	100
688.5067	2	0	120	11.3476	100
699.021	2	0	120	11.68085	100
710.0473	2	0	120	12.37189	100

Table S 13**. Settings of the third 20 windows of the MS scan for DIA analysis**

m/z	z	t start (min)	t stop (min)	Isolation Window (m/z)	Normalized AGC Target (%)
721.2844	2	0	120	12.10229	100
732.6165	2	0	120	12.56189	100
744.5228	2	0	120	13.25073	100
757.0234	2	0	120	13.75037	100
770.527	2	0	120	15.25696	100
784.5499	2	0	120	14.78882	100
798.5827	2	0	120	15.27673	100
813.0505	2	0	120	15.65894	100
827.7426	2	0	120	15.7251	100
843.2513	2	0	120	17.2923	100
859.7603	2	0	120	17.72589	100
877.4602	2	0	120	19.67383	100
897.1549	2	0	120	21.71552	100
918.4784	2	0	120	22.93158	100
941.9666	2	0	120	26.04468	100
968.4867	2	0	120	28.99554	100
1000.831	2	0	120	37.69305	100
1041.069	2	0	120	44.78284	100
1089.265	2	0	120	53.60974	100
1157.785	2	0	120	85.42993	100

Results

SDS PAGE indicated that C_4 species had lower abundances of ribulose-1,5-bisphosphate carboxylase-oxygenase (RubisCO) and higher abundances of PPDK and PEPC than C_3 and C_3-C_4 intermediate species (Fig. S 28). On average, 6,469 proteins (from 6,205 in Fram to 6,640 in Ftri) were detected in each species. For the following analysis, only proteins detected with a unique ID in at least 15 samples were retained, which resulted in 4,908 proteins. We found that the protein abundances of 4,908 proteins across six replicates from a same species showed higher correlations than those between species (Fig. S 29), implying the reliability of sampling and protein quantifications. All proteins were clustered into 10 patterns according to their profiles among the five *Flaveria* species using Kmeans (Fig. S 30). Interestingly, in line with the Kmeans analysis of transcript abundances, C_4 genes were primarily classified into one single pattern, in which the C_4 species Ftri showed the highest abundances. Similarly, photorespiratory genes were mainly clustered into one pattern, in which C_4 species had the lowest abundances (Fig. S 30). C_4 genes had higher protein abundances in the C_4 species Ftri than in other species (Fig. S 31), whereas photorespiratory genes had lower protein abundances in the C_4 species then other species (Fig. S 32), consistent with previous reports [31].

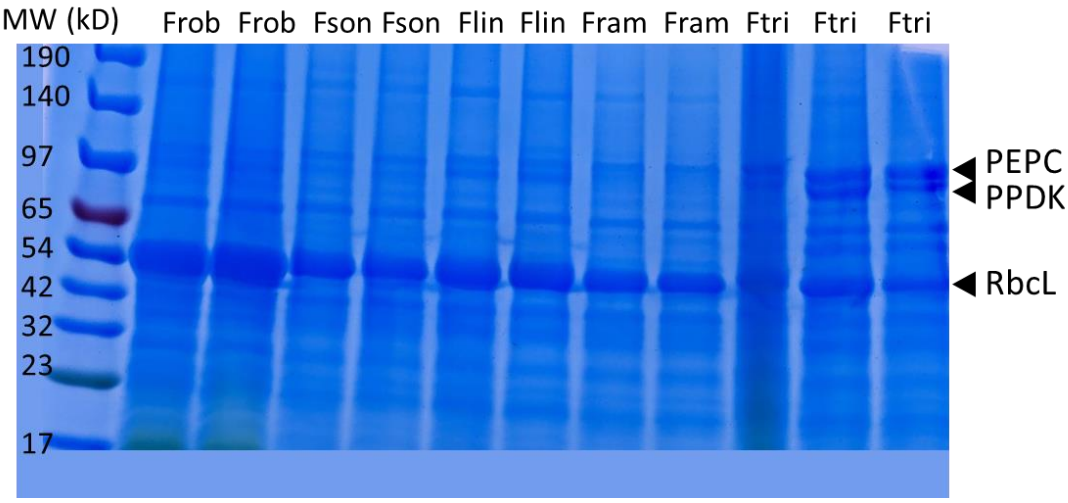


Fig. S 28**. SDS PAGE of the total proteins of *Flaveria* leaves**

Total proteins were isolated from mature leaves and separated with 12% SDS PAGE on the basis of same leaf area. Two or three replicates were run for each species. (Abbreviations: RubisCO: ribulose-1,5-bisphosphate carboxylase-oxygenase, PEPC: phospho*enol*pyruvate carboxylase; PPDK, pyruvate/orthophosphate dikinase.)

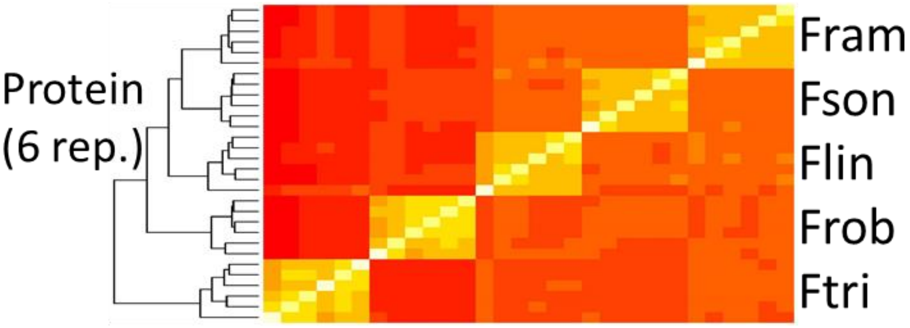


Fig. S 29. **Pearson correlation of protein datasets of *Flaveria* species in protein abundances**

Heatmap showing protein abundances in log2 transformed label free quantification (LFQ) for the five *Flaveria* species. Note that six biological replicates of the same species are clustered as one branch.

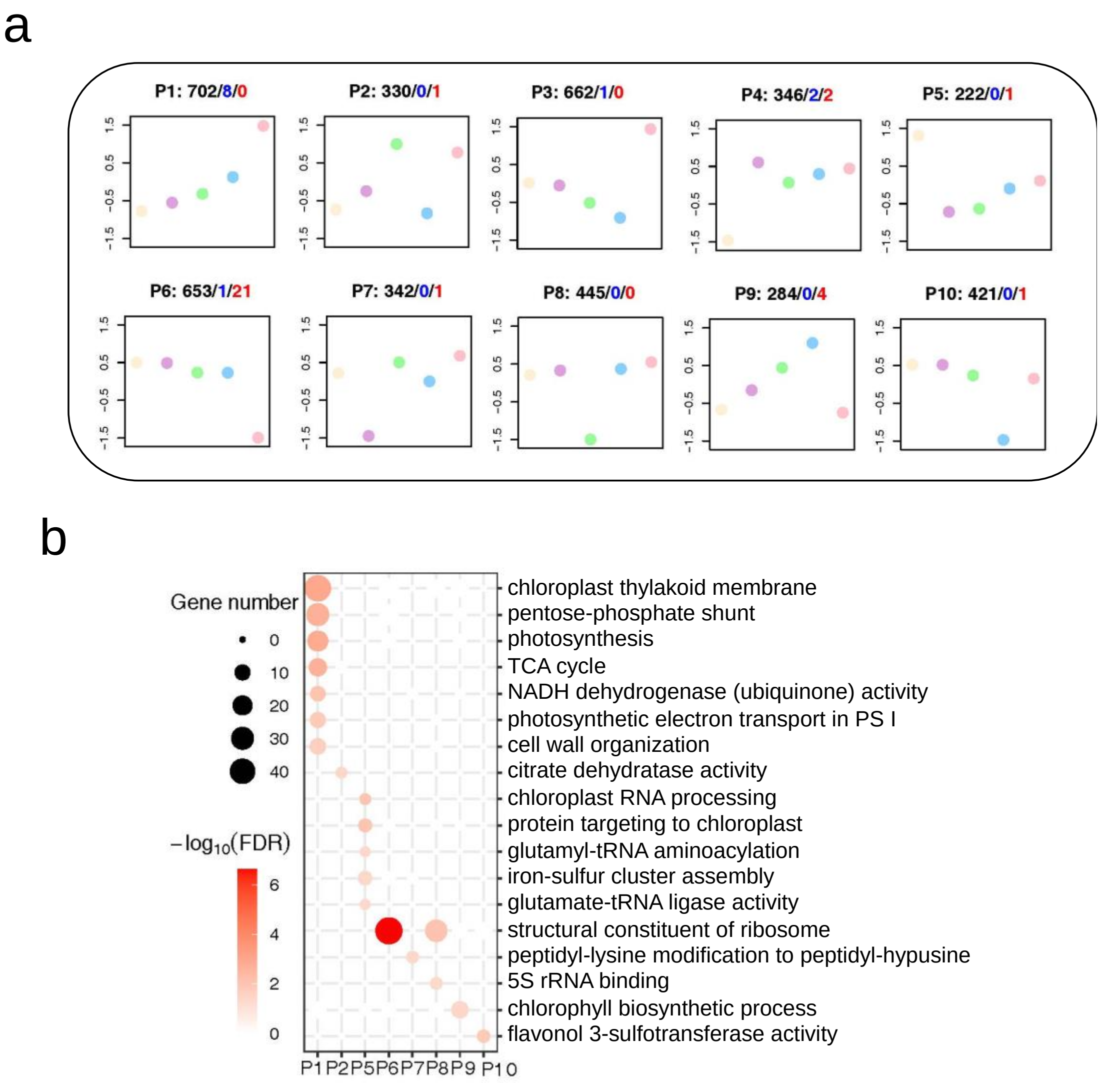


Fig. S 30. **Ten evolutionary patterns of the detected genes based on protein abundances**

(a) Genes were clustered into 10 patterns according to their protein profiles within the five

Flaveria species. Genes were clustered using Kmeans based on the mean protein abundances in label free quantification (LFQ) from six biological replicates for each species. The numbers on the top of each pattern represent: black for all genes, blue for C_4 genes and red for photorespiratory genes. (b) Enriched gene ontology (GO) terms of genes from each pattern. Enriched GO terms were calculated using *Fisher's* test with a P value threshold of 0.01 adjusted applying Benjamini & Hochberg method.

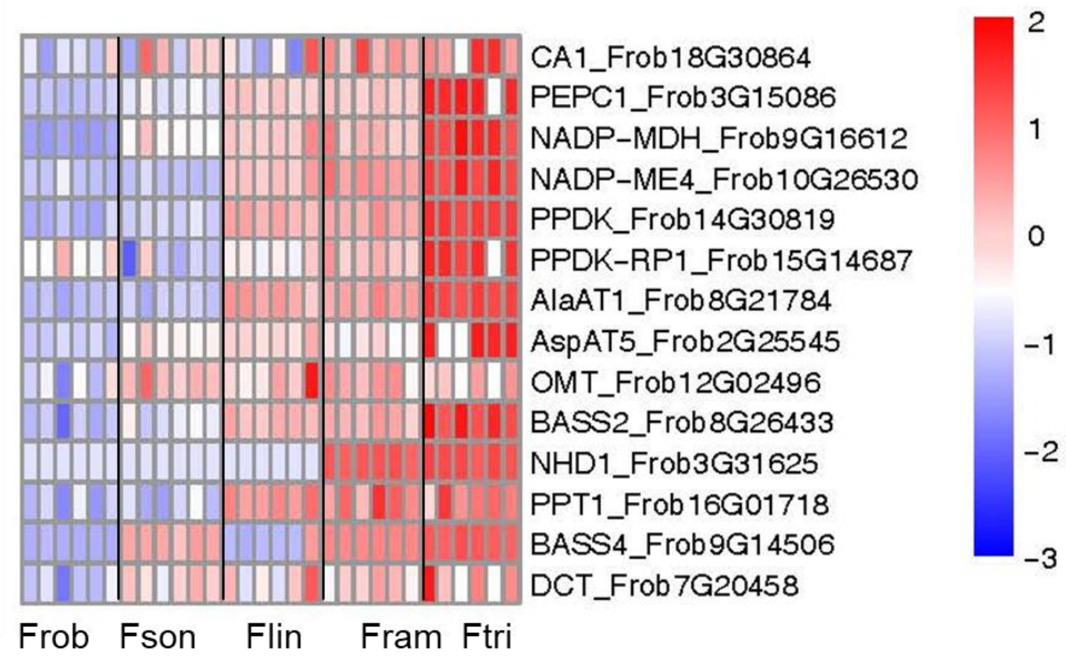


Fig. S 31**. Protein abundances of** C_4 **genes**

Heatmap showing the log2 transformed protein abundances in label free quantification (LFQ). (Abbreviations: CA1: carbonic anhydrase 1; PEPC1: phosphoenolpyruvate carboxylase 1; NADP-MDH: NADP-dependent malate dehydrogenase; NADP-ME4: NADP-dependent malic enzyme 4; PPDK: pyruvate/orthophosphate dikinase; PPDK-RP, PPDK regulatory protein; AlaAT1, alanine aminotransferase 1; AspAT5, aspartate aminotransferase 5; OMT, oxaloacetate/malate transporter, or dicarboxylate transporter 1 (DiT1); BASS2, bile acid sodium symporter 2; NHD1, sodium: hydrogen antiporter 1; PPT1, phosphate/phosphoenolpyruvate translocator 1; BASS4, bile acid sodium symporter 4, and DCT: dicarboxylate transporter 2.1 (DiT2.1))

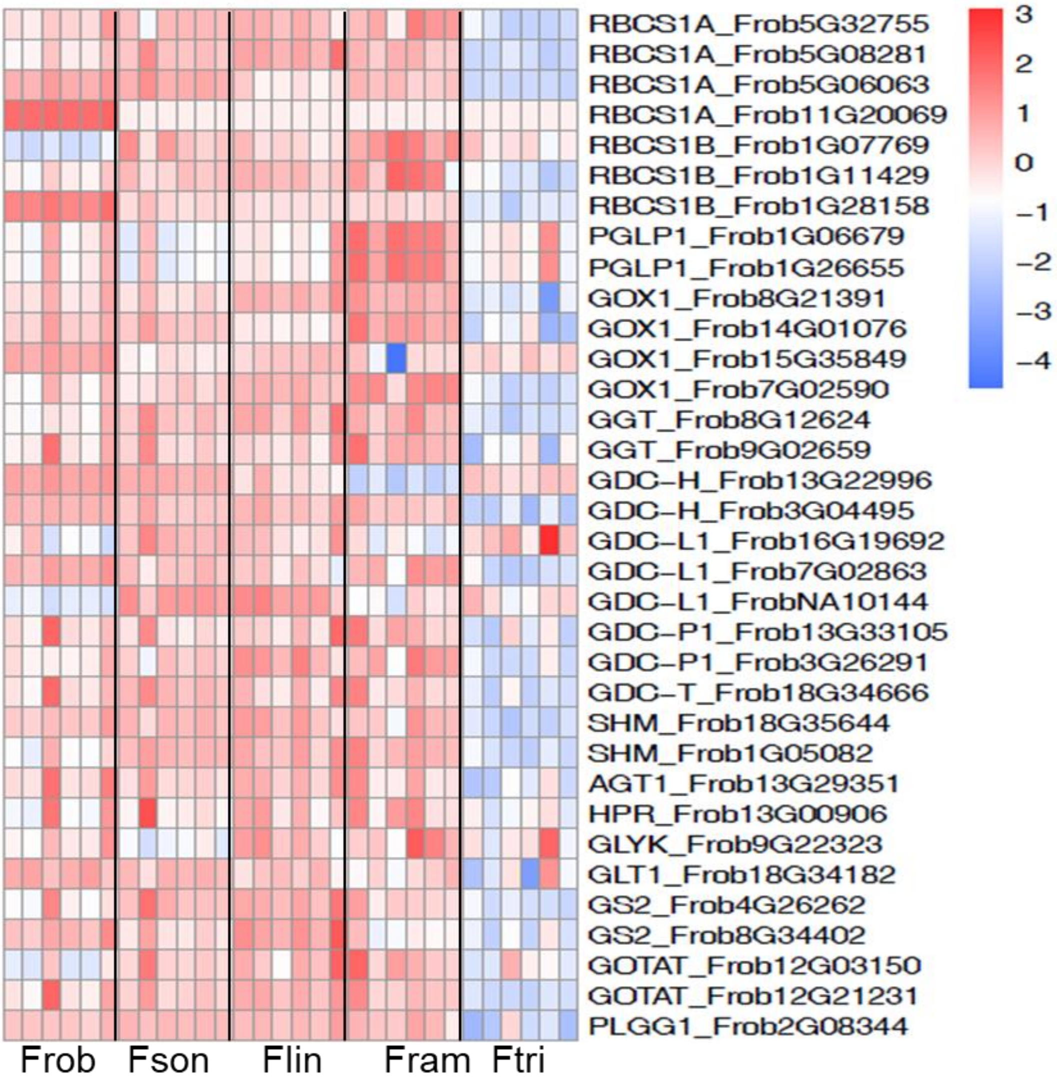


Fig. S 32**. Protein abundance of photorespiratory genes**

Heatmap showing the log 2 transformed protein abundance in label free quantification (LFQ). (Abbreviations: RBCS1A: Ribulose-1,5-bisphosphat-carboxylase/-oxygenase (RubisCO) small subunit 1 A/B; RBSC1B: RubisCO small subunit B; PGLP1: 2-phosphoglycerate phosphatase1, GOX: glycolate oxidase; GDC-H: GDC-L, glycine decarboxylase complex (GDC) subunit L, GDC-P: GDC subunit P; GDC-T: GDC subunit T; SHM: serine hydroxymethyltransferase; AGT1: serine glyoxylate aminotransferase, HPR: hydroxypyruvate reductase ; GLYK: D-glycerate 3-kinase; GLT1: glutamate synthase 1; GS2: glutamine synthase 2; GOTAT: glutamine oxoglutarate aminotransferase; PLGG: chloroplastidic glycolate/glycerate transporter.)

15. Comparison of protein-to-transcript ratio

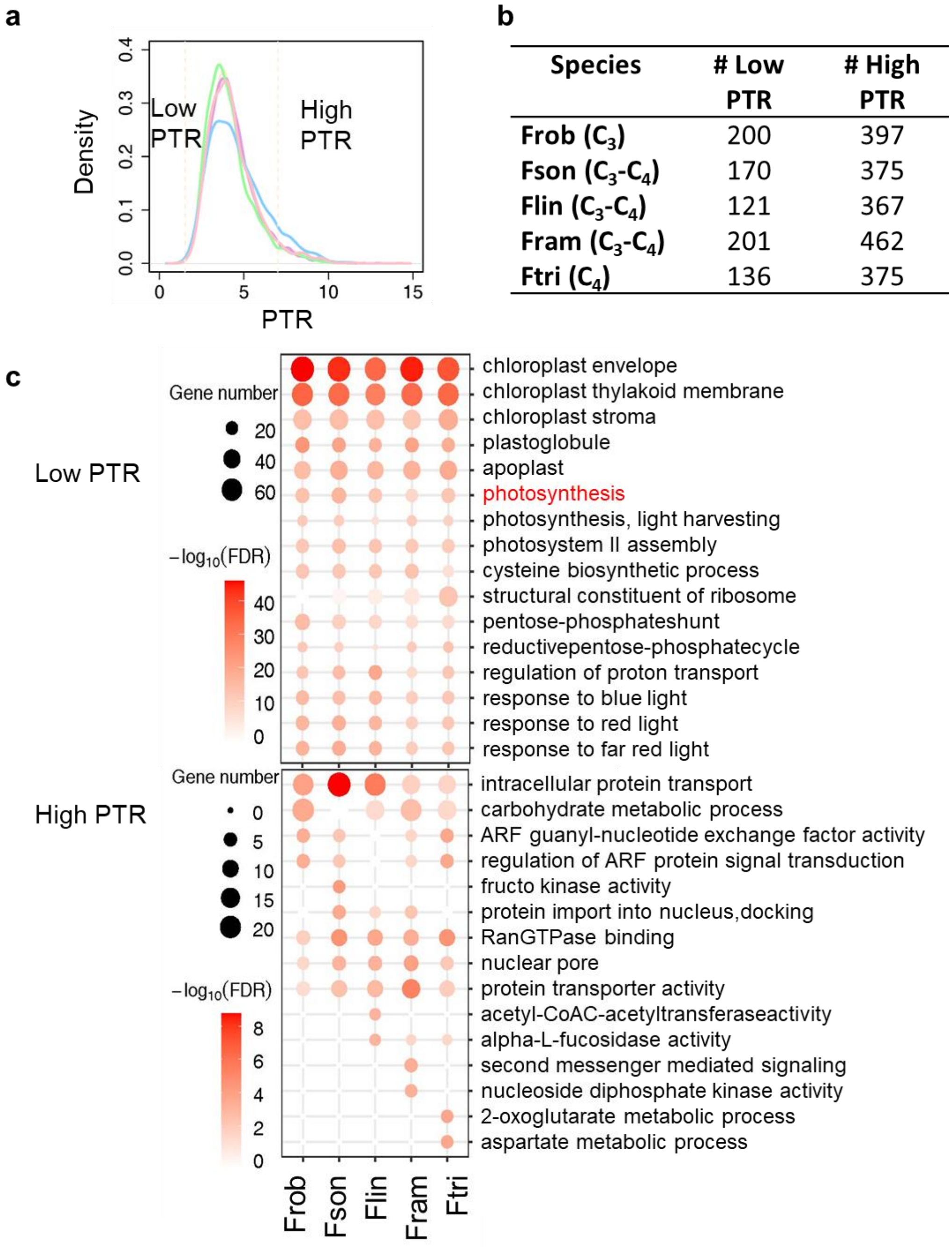
Methods

We compared the protein-to-mRNA ratio (PTR) between genes across five

Flaveria species. Low PTR genes and high PTR genes were defined as genes with PTR values less than the mean PTR minus standard deviation (SD) and higher than the mean PTR plus SD, respectively. The remaining genes were defined as moderate PTR genes. Enriched gene ontology (GO) terms were calculated using *Fisher's* test as mentioned above. The protein abundances and transcript abundances of *Arabidopsis thaliana* (Atha) were inferred from a previously study [32]. To compare PTR between photosynthetic genes and the remaining genes, photosynthetic genes were defined as those with gene ontology of GO:0015979.

Results

An average of 166 low PTR genes (from 121 to 201) and 395 high PTR genes (from 375 to 462) were obtained from the five species (Fig. S 33 a). The low PTR genes were enriched in gene ontology (GO) of photosynthesis and photosynthesis related GO terms, including chloroplast and, PSII (Fig. S 33 c). Additionally, photosynthetic genes showed a trend towards lower PTR (Fig. S 34), consistent with an early study in *Arabidopsis thaliana* (Atha), which showed that photosynthesis related genes had significantly lower PTRs than other genes in photosynthetic functional leaf tissues [32], but not in old leaves (Fig. S 35) or other tissues [32].



Species	# Low PTR	# High PTR
Frob (C_3)	200	397
Fson (C_3-C_4)	170	375
Flin (C_3-C_4)	121	367
Fram (C_3-C_4)	201	462
Ftri (C_4)	136	375

Fig. S 33**. High and low PTR genes and their enriched functions in five *Flaveria* species** (A) The protein-to-mRNA ratio (PTR) distribution of genes from the five *Flaveria* species. High PTR and low PTR genes are defined as genes with PTR higher than the mean plus one standard deviation (SD) and with PTR values lower than the mean minus one SD respectively. (A) the number of low and high PTR genes for each species. (C) Enriched gene ontology of low PTR and high PTR genes, which were calculated using *Fisher*'s test with a *P*-value threshold of 0.001 (Benjamini & Hochberg adjusted). (Abbreviation: PTR: protein-to-mRNA ratio.)

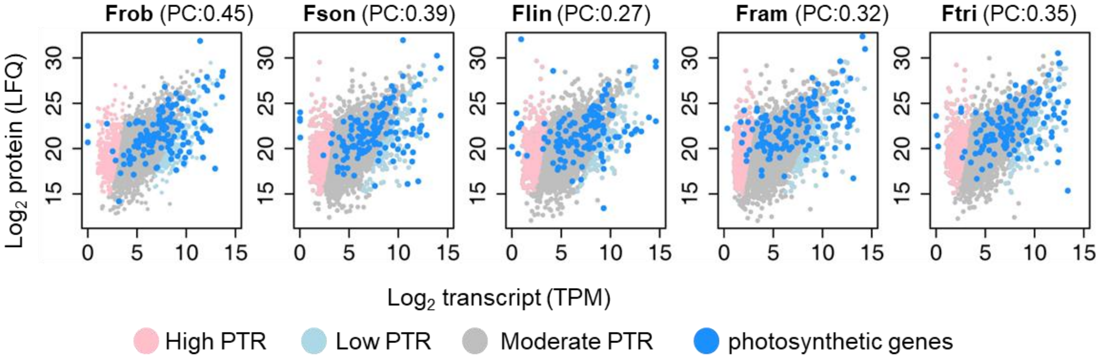


Fig. S 34. **Scatter plots of protein versus transcript abundance in photosynthetic genes**

Pearson correlation (PC) between protein abundance and transcript abundance shown in the parentheses on the top of each panel. Note that photosynthetic genes usually have either moderate or low PTR compared to non-photosynthetic genes in the five *Flaveria* species. (Abbreviations: PTR: protein-to-mRNA ratio; PS. Photosynthesis.)

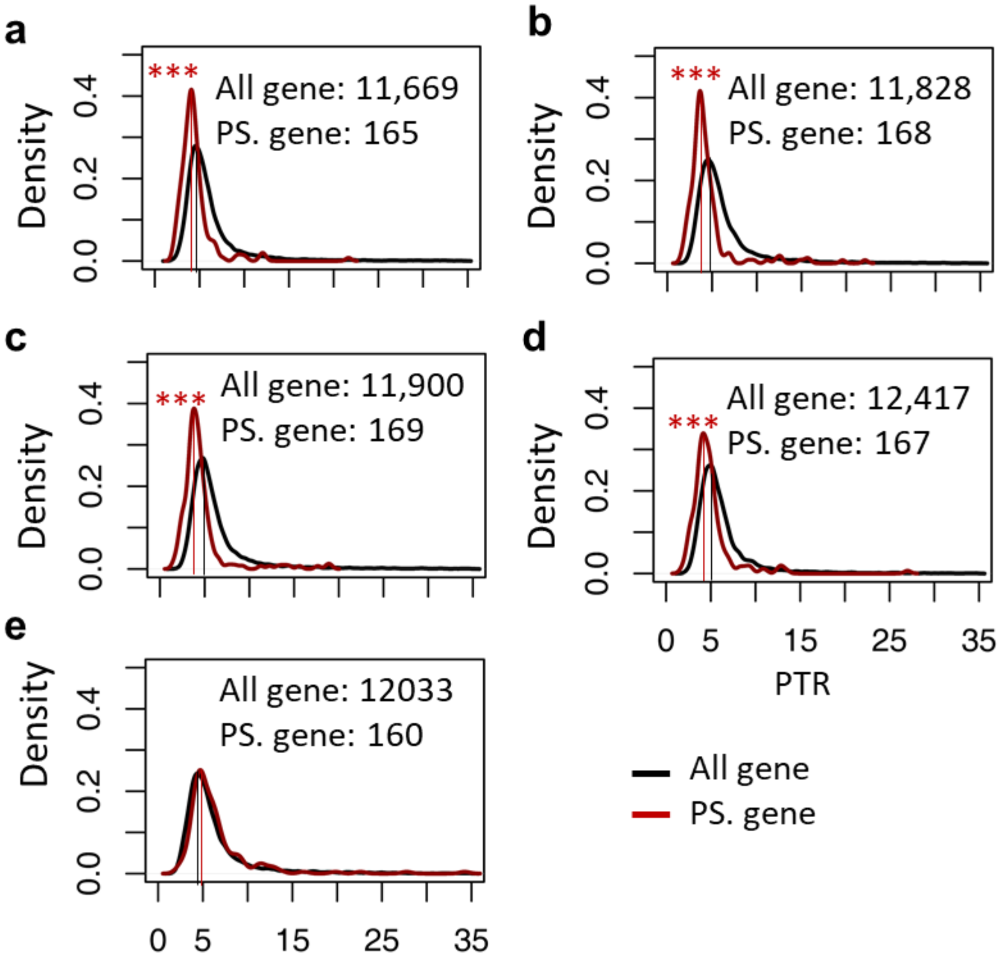


Fig. S 35. **Photosynthetic genes have lower PTRs than other genes in photosynthetic active leaves of *Arabidopsis thaliana***

The distribution of PTRs of all detected genes and photosynthetic related genes in (a) 1st cauline leaf, (b) distal part of rosette leaf 7, (c) proximal part of rosette leaf 7, (d) petiole of rosette leaf 7, and (d) senescent leaf. Photosynthetic genes have annotated gene ontology of GO:0015979. The number of all detected genes and photosynthetic related genes in each leaf sample are showed within the panel. Figures were drawn based on data in supplementary data 4 from a prior study [32] . (Abbreviations: PTR: protein-to-mRNA ratio; PS. Photosynthesis.)

References

1 Li, S. F. *et al.* The landscape of transposable elements and satellite DNAs in the genome of a dioecious plant spinach (Spinacia oleracea L.). *Mob DNA* **10**, 3, doi:10.1186/s13100-019-0147-6 (2019).

2 Seppey, M., Manni, M. & Zdobnov, E. M. BUSCO: Assessing Genome Assembly and Annotation Completeness. *Methods Mol Biol* **1962**, 227-245, doi:10.1007/978-1-4939-9173-0_14 (2019).

3 Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323, doi:10.1186/1471-2105-12-323 (2011).

4 Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21, doi:10.1093/bioinformatics/bts635 (2013).

5 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**, 357-U354, doi:10.1038/Nmeth.1923 (2012).

6 Taniguchi, Y. Y. *et al.* Dynamic changes of genome sizes and gradual gain of cell-specific distribution of C4 enzymes during C4 evolution in genus Flaveria. *Plant Genome*, e20095, doi:10.1002/tpg2.20095 (2021).

7 Min, X. J., Butler, G., Storms, R. & Tsang, A. OrfPredictor: predicting protein-coding regions in EST-derived sequences. *Nucleic Acids Res* **33**, W677-680, doi:10.1093/nar/gki394 (2005).

8 Camacho, C. *et al.* BLAST plus : architecture and applications. *BMC Bioinformatics* **10**, doi:https://doi.org/10.1186/1471-2105-10-421 (2009).

9 Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol* **20**, 238, doi:10.1186/s13059-019-1832-y (2019).

10 Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**, 1792-1797, doi:10.1093/nar/gkh340 (2004).

11 Stamatakis, A. RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688-2690, doi:10.1093/bioinformatics/btl446 (2006).

12 Lyu, M. J. *et al.* What Matters for C_4 Transporters: Evolutionary Changes of Phosphoenolpyruvate Transporter for C4 Photosynthesis. *Front Plant Sci* **11**, 935, doi:10.3389/fpls.2020.00935 (2020).

13 Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Research* **35**, W265-W268, doi:10.1093/nar/gkm286 (2007).

14 Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics* **9**, 18, doi:10.1186/1471-2105-9-18 (2008).

15 Ou, S. & Jiang, N. LTR_retriever: A Highly Accurate and Sensitive Program for Identification of Long Terminal Repeat Retrotransposons. *Plant Physiol* **176**, 1410-1422, doi:10.1104/pp.17.01310 (2018).

16 Kaessmann, H., Vinckenbosch, N. & Long, M. RNA-based gene duplication: mechanistic and evolutionary insights. *Nat Rev Genet* **10**, 19-31, doi:10.1038/nrg2487 (2009).

17 Tan, S. *et al.* DNA transposons mediate duplications via transposition-independent and -dependent mechanisms in metazoans. *Nat Commun* **12**, 4280, doi:10.1038/s41467-021-24585-

9 (2021).

18 Xiao, H., Jiang, N., Schaffner, E., Stockinger, E. J. & van der Knaap, E. A retrotransposon-mediated gene duplication underlies morphological variation of tomato fruit. *Science* **319**, 1527-1530, doi:10.1126/science.1153040 (2008).

19 Tan, S. J. *et al.* LTR-mediated retroposition as a mechanism of RNA-based duplication in metazoans. *Genome Research* **26**, 1663-1675, doi:10.1101/gr.204925.116 (2016).

20 Zhu, Z., Tan, S., Zhang, Y. & Zhang, Y. E. LINE-1-like retrotransposons contribute to RNA-based gene duplication in dicots. *Sci Rep* **6**, 24755, doi:10.1038/srep24755 (2016).

21 Hu, Y. *et al.* Rapid Genome Evolution and Adaptation of Thlaspi arvense Mediated by Recurrent RNA-Based and Tandem Gene Duplications. *Front Plant Sci* **12**, 772655, doi:10.3389/fpls.2021.772655 (2021).

22 Akyildiz, M. *et al.* Evolution and function of a cis-regulatory module for mesophyll-specific gene expression in the C_4 dicot Flaveria trinervia. *Plant Cell* **19**, 3391-3402, doi:10.1105/tpc.107.053322 (2007).

23 Marand, A. P., Chen, Z. L., Gallavotti, A. & Schmitz, R. J. A cis-regulatory atlas in maize at single-cell resolution. *Cell* **184**, 3041-+, doi:10.1016/j.cell.2021.04.014 (2021).

24 Burgess, S. J. *et al.* Genome-Wide Transcription Factor Binding in Leaves from C_3 and C_4 Grasses. *Plant Cell* **31**, 2297-2314, doi:10.1105/tpc.19.00078 (2019).

25 Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38**, 576-589, doi:10.1016/j.molcel.2010.05.004 (2010).

26 Lyu, M. J. *et al.* Evolution of gene regulatory network of C_4 photosynthesis in the genus Flaveria reveals the evolutionary status of C_3-C_4 intermediate species. *Plant Commun* **4**, 100426, doi:10.1016/j.xplc.2022.100426 (2023).

27 Zheng, G. *et al.* CMIP: a software package capable of reconstructing genome-wide regulatory networks using gene expression data. *BMC Bioinformatics* **17**, 535, doi:10.1186/s12859-016-1324-y (2016).

28 Zhang, X. J. *et al.* Inferring gene regulatory networks from gene expression data by path consistency algorithm based on conditional mutual information. *Bioinformatics* **28**, 98-104, doi:10.1093/bioinformatics/btr626 (2012).

29 Reyna-Llorens, I. & Hibberd, J. M. Recruitment of pre-existing networks during the evolution of C_4 photosynthesis. *Philos Trans R Soc Lond B Biol Sci* **372**, doi:10.1098/rstb.2016.0386 (2017).

30 Zhu, M.-J. A. L. J. E. F. C. G. C. X.-G. Evolution of co-regulatory network of C_4 metabolic genes and TFs in the genus Flaveria: go anear or away in the intermediate species? *BioRxiv*, doi:10.1101/2020.10.02.324558 (2020).

31 Mallmann, J. *et al.* The role of photorespiration during the evolution of C_4 photosynthesis in the genus Flaveria. *Elife* **3**, e02478, doi:10.7554/eLife.02478 (2014).

32 Mergner, J. *et al.* Mass-spectrometry-based draft of the Arabidopsis proteome. *Nature* **579**, 409-414, doi:10.1038/s41586-020-2094-2 (2020).