Evolution of co-regulatory network of C₄ metabolic genes and TFs in the genus

Flaveria: go anear or away in the intermediate species?

Ming-Ju Amy Lyu¹, Jemaa Essemine¹, Faming Chen¹, Genyun Chen¹, Xin-Guang Zhu¹

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

#equal contributions

Author for Correspondence
Xin-Guang Zhu, 300 Fenglin Road, Xuhui Area, Shanghai, China, 200032. zhuxg@cemps.ac.cn

Fax: 86-21-54920451
Tel: 86-21-54920486

Abstract

C₄ photosynthesis evolved from the ancestral C₃ photosynthesis by recruiting pre-existing genes to fulfill new functions. The enzymes and transporters required for the C₄ photosynthesis have been intensively studied; however, the transcription factors (TFs) regulating these C₄ metabolic genes are not well understood. In particular, how the TF regulatory network of C₄ metabolic genes was rewired during the evolution is unclear.

Here, we constructed TFs co-regulatory networks for core C₄ metabolic genes (C₄GRN) for four evolutionarily closely related species from the genus Flaveria, which represent four different evolutionary stages of the C₄ photosynthesis, namely, C₃, type I C₃-C₄, type II C₃-C₄ and C₄. Our results show that more than half of the co-regulations of TFs and C₄ core metabolic genes were species specific. The counterparts of C₄ genes in C₃ species were already co-regulated with the photosynthesis-related genes; whereas the required TFs for the C₄ photosynthesis were recruited later. The type I C₃-C₄ species recruited 40% of C₄ required TFs which co-regulated all core C₄ metabolic genes but PEPC; nevertheless, the type II C₃-C₄ species took on a high divergent C₄GRN with C₄ species itself. In C₄ species, PEPC and PPDK-RP possessed much more co-regulated TFs than other C₃ metabolic genes. This study provides for the first time the TFs profiles of the C₄ metabolic genes in species with different photosynthetic types and reveal the dynamic of C₃ genes-TFs co-regulations along the evolutionary process, providing thereby new insights into the evolution of C₄ photosynthesis.

Key words: C₄ photosynthesis, Evolution, Flaveria, GRN, TFs
Introduction

C₄ photosynthesis evolved from the ancestral C₃ type (Sage 2004). In dual cell C₄ species, C₄ photosynthesis compartmentalizes the CO₂ fixation in two cells, where CO₂ is initially converted into a four-carbon acid by phosphoenolpyruvate carboxylase (PEPC), which occurs mainly in the mesophyll cell (MC). The four-carbon acid diffuses then to nearby the bundle cell (BSC), where CO₂ is subsequently released to be eventually fixed by the Calvin-Benson cycle (Hatch 1987). This effective synergistic collaboration between the two cells generates a CO₂ concentrating mechanism (CCM), which results in a high CO₂ concentration around the ribulose-1,5-bisphosphate carboxylase/oxygenase (Rubisco) in the BSC, decreasing thereby the photorespiration rate and leading ultimately to high light, water and nitrogen use efficiencies (Zhu, et al. 2008; Vogan and Sage 2011). The high C₄ photosynthesis efficiency makes it an ideal target and promising to be engineered into C₃ crops for the purpose of increasing the crop yield and productivity (Hibberd, et al. 2008; Sage and Zhu 2011).

All the involved enzymes and transporters in the C₄ photosynthesis are recruited from pre-existing genes of the ancestral C₃ species (Christin, et al. 2013; Moreno-Villena, et al. 2018). So far, the genes encoding the enzymes and transporters in the C₄ core metabolism are well studied (Hatch 1987). The evolution of the C₄ genes and transporters has also been well documented (Christin, et al. 2009; Williams, et al. 2012; Christin, et al. 2013; Moreno-Villena, et al. 2018). However, the regulatory mechanism controlling the C₄ genes remains largely unknown (Hibberd and Covshoff 2010; Schluter and Weber 2020) and also its (regulatory mechanism) evolution constitutes a puzzling enigma. A recent study comparing for the first time the genome-wide transcriptional factor binding sites in three grass C₄ and one grass C₃ species provides a TF binding repertoire in C₄ and C₃ grass species, from which a landscape of TF families associated with those motifs are uncovered (Burgess, et al. 2019). Nevertheless, certain TFs regulating the C₄ genes, especially the regulatory network regrouping the C₄ genes and TFs co-regulations still remains poorly understood (Schluter and Weber 2020). It has been well documented that the cis-elements of C₄ genes exist in their orthologs in C₃ plants (Brown, et al. 2011; Kajala, et al. 2012; Burgess, et al. 2016); moreover, the cis-elements of C₄ genes controlling the cell specificity could be recognized by the TFs in the C₃ species and result in same cell specificity as C₄ species (Brown, et al. 2011; Gorska, et al. 2019; Gupta, et al. 2020). However, it remains unclear to what extent the regulation of C₄ genes required for C₄ photosynthesis could be present in the C₃ species, especially in intermediate species.

Among the model species used to study the C₄ evolution, the genus Flaveria stands out since it comprises species with different photosynthetic types (Powell 1978; McKown, et al. 2005), in which the C₃ species had evolved within 5 million years ago (Christin, et al. 2011). So far, the metabolic and anatomical features of C₃, type I C₃-C₄, type II C₃-C₄, C₄-like and C₄ species in this genus have been extensively studied during the last decade (Sage 2004; McKown and Dengler 2007; Sage, et al. 2012; Mallmann, et al. 2014). The emergence of the type I C₃-C₄ is considered as prerequisite for the C₄ evolution since the formation of C₂ CCM (also known as photorespiratory carbon pump) creates an obvious input of amino residues from MC
into BSC, which requires the rewiring of many C₄ photosynthesis genes (Mallmann, et al. 2014). The type II
C₃-C₄ species gained an increased C₄-ness features, e.g., about 42% CO₂ was initially fixed by PEPC
(Rumpho, et al. 1984). It is worthy to note that both C₄-like and C₃ species have fully functional C₄ metabolic
pathways with the former (C₄-like) still has small fraction of CO₂ being initially fixed by Rubisco (Cheng,
et al. 1988; Dmoore, et al. 1989). Thus, it represents an optimization from the C₄-like to the C₃ stage (Sage,
et al. 2012). Though, an extensive knowledge has been accumulated on the sequences of metabolic changes
in the different Flaveria species during the evolution of C₄ photosynthesis, the knowledge about the molecular
mechanisms underlying these changes was limited and was gained only based on a limited number of genes
such as carbonic anhydrase (CA) (Gowik, et al. 2017), PEPC (Gowik, et al. 2004; Akyildiz, et al. 2007), and

The changes in the genetic regulatory mechanisms usually underlie the progressive changes in the
adaptation and evolution of biological traits (Thompson, et al. 2015). Therefore, elucidation of the changes
in genetic regulatory mechanisms during C₄ evolution is needed to map out the molecular mechanism of C₄
evolution. The construction of a gene regulatory network (GRN) based on the transcriptomic data constitutes
an effective, reliable and powerful approach to systematically decipher the regulatory mechanism of genes
either on the whole genome scale (Friedman 2004) or for a defined genes list for such a metabolic pathway
(Gardner, et al. 2003; Zhang, et al. 2012). In our current study, we aimed to construct the genome-wide GRNs
for four Flaveria species belonging to different photosynthetic stages (C₃, type I C₃-C₄, type II C₃-C₄ and C₄).
The GRNs were developed based on de novo generated RNA-seq data derived from experimental conditions
that have previously been reported to affect the C₄ photosynthesis such as low CO₂ (Sage 2001; Li, et al.
2014), high light (Ubierna, et al. 2013) and exogenous abscisic acid (ABA) treatment (Ueno 2001). With
these GRNs, we systematically study the co-regulated TFs profiling with the C₄ core metabolic genes
(C₄GRN) and how the C₄GRN was re-reshuffled along the C₄ photosynthesis evolution. This report provides
a first glance on the gradual changes of GRN for the C₄ photosynthesis-related genes in one genus
categorized by its short history in C₄ photosynthesis evolution and offers new insights about the metabolic
pathways of the intermediate species.

Results

Transcript assembly and annotation of TFs and C₄ genes in Flaveria species

To identify the TFs co-regulating the C₄ genes, we constructed GRNs for the four Flaveria species based
on the next-generation sequencing (NGS) of RNA-seq data. The species investigated herein represent four
deriving photosynthetic types: F. robusta (C₃; thereafter, Fro), F. sonorensis (type I C₃-C₄; thereafter, Fso),
F. ramosissima (type II C₃-C₄; thereafter, Fra) and F. trinervia (C₄; thereafter, Ftr) according to (Edwards
and Ku 1987). Although, they belong and coexist in the same clade of the phylogenetic tree, the four species
represent four different stages of the C₄ evolution (fig. 1A). The RNA-seq datasets were obtained from plants
subjected to treatments that were earlier reported to regulate the expression levels of the C₄ photosynthesis-
related genes such as low CO₂ (Sage 2001; Li, et al. 2014), high light (Ubierna, et al. 2013) and exogenous ABA application (Fischer, et al. 1986; Duarte, et al. 2019). In total, between 22 and 28 RNA-seq datasets were used to construct the GRN for each species (table 1).

Given that none of the Flaveria species has not yet a published genome sequence so far, the gene annotation for Flaveria species was mainly based on the proteins sequences in Arabidopsis thaliana (A. thaliana) published genome database (Gowik, et al. 2011; Mallmann, et al. 2014; Lyu, et al. 2015), which could not effectively distinguish the different paralogs. Hence, in order to identify the C₄ version of C₄ metabolic genes, we conducted full-length transcript sequences across the transcriptome of Fra using Single Molecular Real Time (SMRT) long read isoform sequencing (Iso-seq) of Pacific Biosciences (PacBio). The Iso-seq helps achieve full-length transcripts and effectively segregates the different paralogs. We then used the assembled gene sequences from the Iso-seq for Fra as a reference dataset to annotate the assembled transcripts for the three other species (fig. 1B). In total, 9.8 million Iso-seq long reads (LR) were obtained for Fra. Subsequently, we corrected the LR with next generation RNA-seq short reads (SR), which resulted in 276,323 transcripts (supplementary fig. S1A).

Thereafter, we assembled the transcripts by combining both LR and SR together. Totally, 515,464 transcripts were found; among which 99,565 were predicted to be protein-coding transcripts with an open reading frame (ORF) no less than 300 bp length. After removal of redundant sequences, 34,045 Fra genes were preserved and used later for further analysis in our current study. The Fra genes were annotated by searching their orthologs from Uniprot, which comprises 31,708,455 entries from Uniref50 and 558,898 entries from Swissprot (Supplementary Material SF2). By seeking for the orthologs in the database PlantTFDB (Jin, et al. 2017), 3,021 TFs from 57 families were annotated and characterized in Fra (Supplementary Material SF2). Among them, 252 TFs belong to bLHL, which represents the largest TF family in Fra, and 216 TFs belong to ERF (supplementary fig. S1B). In line with our findings, BLHL and ERF represent also the top two most abundant TF families in many other plants, such as A. thaliana, Nicotiana tabacum, Oryza sativa and Zea mays (Z. mays), as earlier well documented in the PlantTFDB website (http://planttfdb.cbi.pku.edu.cn/).

On one hand, based on the de novo assembled transcripts from the SR for the four species, between 66,234 and 80,154 transcripts were obtained. To confirm the accuracy of assembling, we further corroborated that around 80 to 96.8% of SR can be mapped to assembled transcripts (Supplementary Material SF3). On the other hand, 99% of the de novo assembled transcripts were predicted to be protein-coding ones (transcripts). Thus, around 60 to 71% of these protein-coding transcripts possess orthologs genes in Fra (supplementary table S1).

We then calculated the gene expression based on NGS in Transcripts per kilobase Per Million mapped reads (TPM) (Supplementary Material SF4). The principle component analysis (PCA) of the gene expression profiling revealed that Fra and Fso are different from the remaining two other species, with the first two components exhibiting 40% of total variance. Fro and Fra could not be separated based only on their first
two PCA components (fig. 1C). However, if only these two species were analyzed using PCA, they can be obviously distinguished into two clusters using the first two components, which reflect a total variance of 52% (fig. 1D).

Since most of the C₄ genes belong to multiple-gene families (Moreno-Villena et al. 2018), it is interesting to determine the C₄ version of C₄ genes. Hence, we manually selected the C₄ version of C₄ genes based on two criteria. (1) The C₄ version gene should show higher transcript abundance compared to other paralogs; (2) The C₄ version should exhibit higher transcript abundance in C₄ species than C₃ ones. Predominately, the C₄ version could mainly be selected based on the first criterion. For instance, two paralogs of β-carboxylic anhydrase 1 (βCA1) were annotated in the C₄ species Ftr with a single paralog (Gene00137) having a transcript abundance of around 20,000 TPM and another paralog (Gene23895) having a transcript abundance of around 1,000 TPM. Besides, the Gene00137 showed higher transcript abundance in the C₄ species Ftr than in the C₃ species Fro (supplementary fig. S2A-D). Accordingly, the Gene00137 was selected to be the C₄ version of CA1 in Ftr. The latter (Ftr) uses a typical NADP-dependent malic enzyme (NADP-ME) type C₄ photosynthesis, where only NADP-ME was up-regulated in Ftr compared to Fro, while NAD-dependent malic enzyme (NAD-ME) and PEP carboxykinase (PEPCK) showed comparable transcript abundance in these four species (supplementary fig. S2A-D). Except the enzymes and transporters required for the NADP-ME pathway, we found as well a significantly higher transcript abundance of alanine aminotransferase (AlaAT) in Ftr if compared to Fro; however, aspartate aminotransferase (AspAT) did not show such a pattern between the four species. We, therefore, included this gene (AlaAT) in the list of C₄ core metabolic genes. The AlaAT is actually known to transfer ammonia from BSC to the MC by glutamate/2-oxoglutarate shuttle in order to maintain the nitrogen homeostasis between the two compartments, MC and BSC (Mallmann, et al. 2014). All the C₄ core metabolic genes show higher transcript abundance in the C₄ species Ftr than that in the C₃ species Fro and such patterns are found to be conserved under the four experimental conditions used in this investigation (supplementary fig. S2). Among the different treatments, four weeks low CO₂ treatment shows the substantially largest effect on the C₄ genes expressions for the four species (supplementary fig. S3).

To confirm the accuracy of the transcript quantification via RNA-seq data processing, a qRT-PCR analysis was performed for 14 genes on the same RNA samples used for RNA-seq in plants exposed for four weeks to low CO₂ of 100 ppm (supplementary fig. S4). Thus, we found an averaged Pearson correlation coefficient of 0.78 between transcript abundance from RNA-seq and that (transcript abundance) from qRT-PCR analysis. This shows the reliability and accuracy of our transcript quantification via RNA-seq data processing. The expression patterns of C₄ core metabolic genes obtained from qRT-PCR are consistent with those evaluated using RNA-seq data (fig. 2A and B). Indeed, figure 2B depicts the proposed C₄ core metabolic pathway based on the identified C₄ genes in Flaveria, which including 8 enzymes: CA1, PEPC1, PEPC-K, NADP-dependent malate dehydrogenase (NADP-MDH), NADP-ME4, AlaAT2, pyruvate/orthophosphate dikinase (PPDK) and PPDK regulatory protein 2 (PPDK-RP2), and 5 transporters: dicarboxylate transporter
1 (DiT1), dicarboxylate transporter 2 (DiT2), bile acid sodium symporter 2 (BASS2), sodium: hydrogen antiporter 1 (NHD1) and phosphate/phosphoenopyruvate translocator 1 (PPT1).

The properties of C₄ GRNs in four Flaveria species

For this purpose, we constructed a genome-wide GRN based on the RNA-seq data for each species. We further built a “C₄ GRN” which retains only the C₄ genes and their positively co-regulated TFs. We termed this proposed network as “C₄ GRN” and designated the TFs involved in C₄ GRN as “C₄ TFs” (Supplementary Material SF5). The genome-wide GRN of type I C₃-C₄ species Fso contains the minimum genes (14,677 genes) and the maximum interactions (7,841,065 interactions). In contrast, the C₄ species Ftr contains 14,761 genes and 5,769,379 interactions (fig. 3A). Similarly, the C₄ GRN of Fso includes 220 TFs, followed by the networks of Ftr and Fso and Fra with 196, 185 and 114 TFs, respectively (fig. 3A). In the same regard, more than half of the identified C₄ TFs are found to be specific for each species. For instance, 57.9% C₄ TFs of Fra is not shared with any other species and the percentage of species-specific C₄ TFs in Fso is 69.2% (fig. 3B).

In total, 36.2% C₄ TFs in Ftr overlapped with those of at least one of the three other species. The C₄ GRN of Fso exhibits 37 TFs overlapped with Ftr, displaying the highest overlap degree with C₄ species compared to Fro and Fra (fig. 3C). However, the co-regulated C₄ genes of these overlapped C₄ TFs changed in these different Flaveria species. Five C₄ genes-TFs co-regulations are common between Fro and Ftr, involving five C₄ genes, which are NHD1, NADP-MDH, DiT2, BASS2 and PEPC-k. The same number of C₄ genes-TFs co-regulations exists between Fso and Ftr, among them three involving the following genes: PPDK-RP, DiT1 and NHD1. Notably, Fra shares seven C₄ genes-TFs co-regulations with Ftr with four engaging PEPC1 (fig. 3C). The other three co-regulations targeted three other genes including DiT1, NHD1 and NADP-MDH.

In the C₄ GRN of Ftr, PPDK-RP shows the greatest number of co-regulated TFs with 63 TFs, followed by PEPC1 with 45 TFs, while CA1 ranks the latest with only one TF (fig. 3C). Interestingly, 74 C₄ TFs (37.8%) in Ftr show the highest transcript abundance in Fso among the four species, even more than the number of C₄ TFs showing the highest transcript abundance in Ftr (62 TFs), which represents 31.6% from the total C₄ TFs of Ftr (fig. 3D).

The most abundant C₄ TFs in the four species is bHLH (fig. 3E). The next most abundant TFs in both Fro and Fso is MYB-related TFs, whereas Fra has NAC family TFs being the second most prevailing TFs; however, GRAS family TFs is the second most abundant, after bHLH, in the Ftr. In fact, 21 C₄ TFs belonging to the GRAS family in Ftr are co-regulated with 9 over 13 C₄ genes (fig. 3F).

Therefore, compared to C₃ and both types of the intermediate species, the C₄ species Ftr recruited a large number of TFs, which are co-regulated with PEPC1 or PPDK-RP. Simultaneously, several TFs from the GRAS family were also recruited. Fso shares the largest number of C₄ TFs with the C₄ species, while Fra shares the maximum C₄ genes-TFs co-regulations with the C₄ species, where four of them (co-regulations) are associated with PEPC1.
Assigned functions of C₄ TFs and C₄ genes

To gain insights on the functions of C₄ TFs in each species reported herein, we firstly investigated the enriched functions of C₄ TFs. The obtained results show that C₄ TFs from the four species consistently are enriched in glycolysis (SI6). The C₄ TFs from Fso and Fra show more enriched GO functions than Fro and Ftr. In Fso, C₄ TFs are also enriched in histone acetyltransferase activity and brassinosteroid mediated signaling pathway (Supplementary Material SF6); whereas, in Fra, C₄ TFs are enriched in functions related to glycine mentalism, i.e., glycine dehydrogenase (decarboxylating) activity and glycine catabolic process. However, None C₄ TFs from all the species exhibit over-represented functions directly related to photosynthesis.

Afterwards, we investigated the related functions of C₄ TFs in each species by examining the functions of their co-regulated genes. To achieve that, we firstly clustered the whole genome-wide GRN into multiple sub-networks (modules) and then studied the biological function of the modules enriched in C₄ TFs (MET; fig. 4A). Specifically, 222, 363, 642 and 295 modules were identified in Fro, Fso, Fra and Ftr, respectively (fig. 4B). We found only a single MET in Fro, Fso and Ftr individually, while two METs were characterized in Fra (Fisher’s test, P < 0.001, BH adjusted); however, uniquely one of the METs in Fra showed over-represented GO functions (fig. 4B).

We further examined the top five enriched GO terms of genes in METs from three aspects, i.e., biological process (BP), cell component (CC), and molecular function (MF). Thus, Fro exhibits a large proportion of GO terms from BP related to the transcriptional and post transcriptional processes, e.g., RNA methylation, tRNA metabolic process and translation. This species (Fro) includes also a single GO term associated with the cellular nitrogen compound metabolic process (fig. 4C). In terms of CC, the enriched GO functions are mainly related to the various cellular compartments, such as mitochondrion, organelle, intracellular, cytoplasm and ribosome (fig. 4C), but not the chloroplast.

Indeed, a substantial overlap exists, regarding the enriched GO terms in METs, between Fso, Fra and Ftr. Hence, many of these overlapped GO terms of BP are involved in photosynthesis, or terms directly related to photosynthesis, such as photosynthesis and thylakoid membrane organization; and some other items indirectly linked to photosynthesis, including carotenoid biosynthetic process, chloroplast relocation, response to red light and to far-red light (fig. 4C). Two other biological processes, i.e., the pentose phosphate shunt and lipid metabolic pathway, are also found among the overlapped GO terms in Fso, Fra and Ftr.

Consistently, almost the whole enriched GO functions in the terms of CC are predominately attributed to the chloroplast term such as chloroplast, chloroplast envelope, chloroplast thylakoid membranes, chloroplast stroma, and thylakoid (fig. 4C).

We thereafter investigated the functions of the C₄ orthologous genes in different species by looking after the over-represented GO terms in modules that include the maximum C₄ orthologous gene (MMC₄). In Fso, the MET and MMC₄ are the same module, which includes five C₄ genes, which are CA1, DiT2, NADP-ME4, NADP-MDH and PPDK. Two MMC₄ were found in Ftr with one coincides with the MET in Ftr. A
considerable fraction of the over-represented GO terms of MMC4 are shared among Fro, Fso and Ftr, and many of them are related to the photosynthesis and chloroplast (fig. 4D). The shared GO terms, associated with BP, between Fro, Fso, and Ftr, include photosynthesis, carotenoid biosynthetic processes, pentose-phosphate shunt, cofactor metabolic process, sulfur metabolism, isopentenyl diphosphate biosynthetic processes, small molecule metabolic pathways, responses to red and far-red light, generation of precursor metabolite and energy. Furthermore, the shared GO terms, related to CC, between the three species (Fro, Fso, Ftr) enlobe thylakoid, chloroplast membrane and envelope. Most of these enriched GO terms in Fro, Fso and Ftr were not shared with Fra (fig. 4D).

Taken together, the orthologs of C3 key metabolic genes in the genus Flaveria were already co-regulated with the photosynthetic genes in C3 species, whereas, the C4 photosynthesis related regulators were widely recruited during the latest C4 evolution stages.

The overlap of C4GRNs between the intermediate and C4 species

Our above analysis shows that many C4TFs are co-expressed with the photosynthesis related genes in two intermediate C3-C4 species (fig. 4C). Given the metabolic and anatomical adjustments between the intermediate species and the C4 ones (species), what are the missing regulatory factors related to C4 photosynthesis in these intermediate C3-C4 species compared to C4 species? On one hand, we found that the 37 C4TFs shared between Fso and Ftr are co-regulated with 12 of the 13 C4 genes (except the PEPC1, fig. 5A). This is consistent with the lower transcript abundance of PEPC1 in Fso (fig. 2B and supplementary fig. S2A-D). On the other hand, in the type II C3-C4 species Fra, 26 C4TFs were found to be common between Fra and Ftr, which are co-expressed with 8 over 13 C4 genes in Fra (fig. 5B). Four among the 26 TFs are co-regulated with PEPC1 (fig. 3C and Supplementary Material SF 5).

A mesophyll expression module 1 (MEM1) located on the PEPC1 promoter region of the C4 species F. bidentis and Ftr was reported to control the mesophyll specific high expression of PEPC1 (Gowik, et al. 2004). We examined, herein, whether the MEM1 is present in the PEPC1 promoter region of Fra. Indeed, the MEM1 is C4 type in Fra, while it is a C3 type in Fro and Fso (fig. 5C and Supplementary Material SF7), which is in consistency with the up-regulation of PEPC1 expression in Fra compared to Fro and Fso (fig. 2B and supplementary fig. S2). The four TFs mentioned above to be co-regulated with PEPC1 in both Fra and Ftr (fig. 5B) are involved or annotated as a negative regulation function (Gene03809), a TF belongs to NF-YA7 family (Gene12432), regulation of drought-responsive genes (Gene17333) and repressor of gibberellins response (Gene18956, supplementary table S2).

The shared 37 TFs between Fso and Ftr are mainly enriched in two GO functions, proton-exporting ATPase activity (phosphorylative mechanism) and other cellular processes (Supplementary Material SF6). However, the shared 26 TFs between Fra and Ftr are enriched in GO functions of glycolysis and chloroplast membrane (Supplementary Material SF6). Compared to Fso, Ftr newly gained 159 TFs and lost 183 TFs (fig. 3B). The newly recruited TFs are enriched in GO functions of glycolysis, chloroplast, nucleus and coenzyme
A metabolic process. Otherwise, the lost TFs (present in Fso but not in Ftr) are enriched in glycolysis, mitochondrion, histone acetyl-transferase activity and isocitrate dehydrogenase (NAD\(^+\)) activity (Supplementary Material SF6). Compared to Fra, Ftr newly gained 170 TFs and lost 88 TFs. The gained TFs are enriched in GO functions of nucleus, coenzyme A metabolic process and hydroxyl-methyl-glutaryl-CoA reductase (NADPH) activity; whereas, the lost TFs (present in Fra but not in Ftr) are essentially enriched in GO items of glycine dehydrogenase (decarboxylating) activity and glycine catabolic process (Supplementary Material SF6). Therefore, C\(_4\)TFs in the two intermediate species as well as C\(_4\) species are involved in divergent functions.

Taken together, it seems that the mechanisms underlying the mesophyll specific high expression of PEPC1 are not yet recruited in Fso. However, in Fra, the mechanisms governing the cell-specific high expression of PEPC1 appears, at least, partially recruited. Nevertheless, five C\(_4\) genes including PPDK, NADP-ME4, PPDK-RP, AlaAT2, PPT1 are exempted from the co-regulation process by the shared C\(_4\) TFs between Fra and Ftr, which is in agreement with the difference between the enriched GO terms in MMC\(_4\) for Fra compared to those for Ftr (fig. 4D).

The evolution of C\(_4\) GRN towards a C\(_4\) photosynthesis stage

In this section, we illustrated how the C\(_4\) GRN of C\(_4\) species gradually evolved during the evolution process. We firstly defined the C\(_4\) required TFs (thereafter, C\(_4\)ReTFs) as C\(_4\)TFs in Ftr that show significantly higher transcript abundance in Ftr than in the C\(_3\) species Fso (\(P < 0.05\) and fold change (FC) \(> 1.5\)). Eventually, we identified 93 C\(_4\)ReTFs, targeting all the 13 C\(_4\) genes in Ftr (fig. 6A).

In Fso, all the shared 37 C\(_4\) TFs with Ftr are included in the 93 C\(_4\)ReTFs (fig. 6B). We classified the modified C\(_4\)ReTFs between Fso and Ftr into two types (fig. 7A). The first one includes the existing C\(_4\)ReTFs in Fso, which exhibited enhanced expressions in Ftr. We found that 10 out of the 37 C\(_4\)ReTFs (fig. 6A and B) displayed higher transcript abundance in Ftr than in Fso (\(P < 0.05\) and FC \(> 1.5\)), with five of them being co-regulated with PEPC1 in Ftr (fig. 6A). The second one (type) encompasses the absent C\(_4\)ReTFs from the C\(_4\)GRN of Fso. In this latter type, we identified 56 C\(_4\)ReTFs (fig. 6A). Besides, new co-regulations were developed between the first type of C\(_4\)ReTFs and the C\(_4\) genes during the transition from type I C\(_3\)-C\(_4\) to C\(_4\) species (fig. 7 and supplementary fig. S5).

Seven of the 37 C\(_4\)ReTFs recruited in Fso are present in the C\(_4\) GRN of Fro and co-regulate with three C\(_4\) core metabolic genes in Fro, i.e., NADP-MDH, BASS2, PEPC-k (fig. 6C). Six over the seven TFs exhibit significantly higher transcript abundance in Fso than in Fro (\(P < 0.05\) and FC \(> 1.5\); fig. 6B). Among the 30 C\(_4\)ReTFs that were newly recruited in Fso, 17 are up-regulated in Fso if compared to C\(_3\) species (\(P < 0.05\) and FC \(> 1.5\); fig. 6B).

Discussion
A comparative analysis of the GRN between species could shed light on the driving force and the mechanisms underlying the adaptation and evolution processes (Thompson, et al. 2015). In this study, we provide a comprehensive survey about the changes of GRN in four species from the genus Flaveria, representing four different stages in C_4 evolution, i.e., Fro (C_3), Fso (type I C_3-C_4), Fra (type II C_3-C_4) and Ftr (C_4) (Edwards and Ku 1987). Considering that the transition from C_3 to C_4 photosynthesis spans a short evolutionarily divergence time (Christin, et al. 2011), the comparison between the GRNs for these four species could shed new light on the evolution of C_4 photosynthesis.

**Fso recruited a large number of C_4ReTFs very likely through reorganizing pre-existing cis-elements**

_Fso_ (type I C_3-C_4) shares the highest number of C_4TFs overlapping with _Ftr_ (C_4), with a number of 37 (fig. 3B), which were predicted to regulate 12 of the 13 C_4 core metabolic genes studied in this work (fig. 5). All the 37 TFs were characterized as C_4ReTFs since they show higher expression levels in the C_4 _Ftr_ than in the C_3 _Fro_ (fig. 6B). Besides, the MET and MMC4 coincide to form the same module in _Fso_, which is a large fraction of the enriched GO functions related to photosynthesis (fig. 4). This is in agreement with the current notion that the metabolism of C_3-C_4 species, which performs C_2 CCM, might represent a metabolic pre-condition for the emergence of C_4 photosynthesis, since many of the involved enzymes and transporters in this poise are common with those in C_4 photosynthesis (Mallmann, et al. 2014). Four of the 37 TFs were annotated to be stress response related (Gene02598 and Gene29705: DBB family, salt stress; Gene04536: ERF family, regulation of expression by stress factors and Gene29301: bHLH family, oxidative stress, Supplementary Material SF8). This is in line with earlier reports that C_2 CCM evolved to cope with conditions favoring photorespiration (Sage, et al. 2012; Lundgren and Christin 2017). The large number of the recruited C_4ReTFs in _Fso_ might be accomplished by recruiting pre-existing cis-elements and/or rewiring pre-existing TFs to gain new functions (Brown, et al. 2011; Reyna-Llorens, et al. 2018).

In this analysis, we found that 30 over the recruited 37 C_4ReTFs are not represented in the C_4GRN of _Fro_, C_3 (fig. 6C). These TFs might be recruited through hiring the already existing cis-elements into C_4 genes. In this regard, it has been reported that the C_4 genes from _Z. mays_, C_4, harbor a number of conserved cis-elements with their counterparts in _Oryza sativa_, C_3 (Xu, et al. 2016). Furthermore, multiple C_4 genes may recruit the same (or similar) cis-elements. For example, PEPC1 and CA3 (named CA1 in this study) harbor MEM1 and MEM1-like motif that determine the MC expression specificity in the C_4 species _F. bidentis_ (Akyildiz, et al. 2007; Gowik, et al. 2017); CA and PPDK share a MEM2 motif in 5’UTR in the C_4 species _Gynandropsis gynandra_ (G. gynandra) which controls the expression specificity in the MC (Williams, et al. 2016). In this context, it has been also reported that cis-elements from the C_4 genes could be recognized by the TFs of C_3 species (Ku, et al. 1999; Nomura, et al. 2000; Fukayama, et al. 2001; Akyildiz, et al. 2007). In the case of NAD-ME, the cis-element controlling the BS specific expression in the C_4 species _G. gynandra_ similarly shows the same cell specificity in the C_3 species _A. thaliana_ (Brown, et al. 2011). Furthermore, a comparable cis-element was recruited into PEPC in three grass species (Gupta, et al. 2020). Altogether, this
shows that a cis-element repository already existing in the ancestral C₃ species could be recruited for the C₄ photosynthesis.

Interestingly, 23 of the 37 C₄ReTFs exhibit higher transcript abundance in Fso than in Fro, suggesting that some modifications in the TFs had also occurred during the evolution. In this context, it has been also reported that the cis-elements from PPDK, CA (Kajala, et al. 2012) and NAD-ME (Brown, et al. 2011) from C₃ species can achieve the C₄-type cell specific accumulation of GUS in C₄ species. Therefore, the TFs in C₄ species might have gained an increased expression or altered their cellular expression pattern to use existing cis-elements and thereby gain new regulatory functions. For instance, it has been previously reported that nine TFs acquired shared MC- or BSC-preferential expression patterns in two divergent C₄ lineages: G. gynandra and Z. mays (Aubry, et al. 2014).

It is worthy to note that, due to the absence of genome sequence and annotation for any Flaveria species so far, it remains difficult to estimate to which extent the cis-elements of the C₄ orthologous genes in Fso were newly recruited and whether the C₄-type expression patterns of the recruited C₄ TFs were established in the same species (Fso).

**Does Fra represent a dead-end for the C₄ evolution?**

The current notion of C₄ evolution in the genus Flaveria is that the C₃-C₄ intermediates have two types, with Fra represents the type II C₃-C₄ intermediate, which shows more C₄-ness than the type I C₃-C₄ species does (McKown and Dengler 2007; Vogan and Sage 2011). Given this notion, it is striking that Fra shows less C₄-like C₄GRN than Fso does, as reflected by the absence of a number of critical C₄ genes in the overlapping C₄GRN between Fra and Ftr (fig. 5A). Furthermore, in Fra, the genes in the MMC₄ display no enriched GO functions in photosynthesis (fig. 4D). Therefore, it raises a question regarding whether Fra constitutes a transitional state from type I C₃-C₄ to C₄ photosynthesis. On one hand, compared to type I C₃-C₄ intermediate, Fra indeed shows increased vein density (McKown and Dengler 2007), an increased cyclic electron transport capacity (Vogan and Sage 2011; Nakamura, et al. 2013), lower CO₂ compensation point (Ku, et al. 1983; Nakamoto, et al. 1983; Rumpho, et al. 1984; Ku, et al. 1991) and higher expression levels for many C₄ genes (Gowik, et al. 2011; Mallmann, et al. 2014). Besides, 42% of CO₂ can initially be fixed into malate and aspartate in Fra (Rumpho, et al. 1984). However, on the other hand, Fra also shows C₃ characteristics for a number of other C₄ related features, such as CO₂ assimilation rate (Ku, et al. 1991), δ¹³C (Gowik, et al. 2011) and the water using efficiency (Vogan and Sage 2011). The C₃-level of the δ¹³C implies that there is no substantial increase in the C₄ flux in the Fra compared to the C₃ Ftr.

However, we did find that four C₄TF and PEPC1 co-regulations are conserved between Fra and Ftr (fig. 5B). Moreover, Fra shows C₄ type MEM1 motif at its promoter region (Fig. 5C). All these are in line with the increased transcript abundance for PEPC1 in Fra compared with that of Fro and Fso (Gowik, et al. 2011; Mallmann, et al. 2014), which implies the increased role of PEPC in fixing CO₂ in this species. However, the fixed CO₂ might not be pumped to the BSC through the malate or aspartate as the typical C₄ metabolic
pathways do. Because if it is the case, an increase in $\delta^{13}$C would be expected, as for the typical C$_4$ species (Farquhar, et al. 1989). Furthermore, neither NADP-MDH, which converts OAA to malate, nor AspAT, which converts OAA to aspartate, shows increased transcript abundance in _Fra_ compared to the C$_3$ _Fro_ (supplementary fig. S2A-D; (Gowik, et al. 2011; Mallmann, et al. 2014). Therefore, the fixed CO$_2$ in the form of OAA might be used to support metabolic reactions other than the C$_4$ cycle.

If there is no increased C$_4$ metabolic flux in _Fra_, so what might underlie the observed enhanced C$_4$-ness in _Fra_? The high C$_4$-ness of _Fra_ most probably resulted from the increased C$_2$ CCM. An increased PEPC1 function would result in a low CO$_2$ concentration in the MC, which could increase the photorespiratory fluxes. Indeed, _Fra_ has relatively higher expression levels for the photorespiratory genes if compared to both C$_3$ and C$_4$ species (Mallmann, et al. 2014). The increased photorespiratory flux in MC could enhance the glycine shuttle from the MC towards the BSC, where glycine would be converted to serine, leading to an accentuated release of the CO$_2$ and ammonium in the BSC and consequently favors the C$_2$ CCM. The increased C$_2$ CCM does not affect $\delta^{13}$C as much as in the case of C$_4$. Furthermore, in _Fra_, the CO$_2$ fixed by PEPC1 might be released into the cytosol through the TCA cycle. In another word, the primary metabolism of _Fra_ may represent an alternative strategy to strengthen the CCM. In this regard, earlier studies showed that the intermediate species _Alloteropsis semialata_ of the grass family (Lundgren, et al. 2016), C$_3$-C$_4$ intermediate species from Moricandia such as _M. arvensis_ and _M. suffruticosa_ (Schluter, et al. 2017), and C$_3$-C$_4$ species _Salsola divaricata_ (Lauterbach, et al. 2017) displayed comparable values of $\delta^{13}$C to that of the C$_3$ species. The metabolic status of the C$_3$-C$_4$ intermediate species deserves further detailed studies.

**Potential existence of a master regulator mediating the C$_4$ photosynthesis**

In the C$_4$ species _Ftr_, many TFs were shown to be co-regulated with multiple C$_4$ genes, *e.g.*, two TFs (Gene24690 and Gene02109) are co-regulated with the same four C$_4$ genes including PPDK-RP, BASS2, DiT2 and PPT1 (fig. 3C and Supplementary Material SF5). Furthermore, GRAS is the second most abundant TF family in the C$_4$GRN of _Ftr_, with 21 GRAS TFs co-regulating 9 over 13 C$_4$ genes (CA1, PEPC1, NADP-ME4, NADP-MDH, PPDK, PPDK-RP, DiT1, DiT2 and NHD1; fig. 3F). Given that the TFs of GRAS family are reported to also control the development of Kranz anatomy, such as Scarecrow (SCR), Shortroot (SHR) (Slewinski 2013; Wang, et al. 2013; Fouracre, et al. 2014) and Golden2-like (GLK) (Wang, et al. 2017), the TFs in the GRAS family might co-regulate the C$_4$ metabolism and anatomy simultaneously. Another significant change in the C$_4$TF network of _Ftr_ compared to the three other Flaveria species is the recruitment of a large number of TFs for PEPC1 and PPDK-RP. This recruitment of TFs regulating PPDK-RP agrees with the crucial and prominent role for the PPDK-RP in maintaining a high efficient C$_4$ metabolism in _Z. mays_ (Hart, et al. 2011). Here we found that TFs from the GRAS family co-regulate these two enzymes (fig. 3F), again suggesting a possibility that members of the GRAS family TFs may represent candidate master regulators for C$_4$ photosynthesis evolution in the genus Flaveria. Therefore, efforts are warranted to identify TFs co-regulating both C$_4$ anatomical and metabolic features to support C$_4$ crop engineering.
Materials and Methods

Plant materials and growth conditions

Seeds of *Fro* (C₃) and *Fra* (C₃-C₄, type II) were provided by Prof. Peter Westhoff (Heinrich Heine University, Germany). Seeds of *Fso* (C₃-C₄, type I) and *Ftr* (C₄) were obtained from Prof. Rowan Sage (University of Toronto, Canada). For low CO₂ experiment, plants were grown in the growth chamber with a photosynthetic photon flux density (PPFD) controlled to be 200 μmol m⁻² s⁻¹, temperature 22 ± 2 °C, 70% relative humidity (RH), and a photoperiod of 16 h light/8 h dark. The CO₂ concentration used for the low CO₂ treatment was 100 ppm and the control CO₂ was around 380 ppm. Plants for ABA treatment were grown in the phytotron of Shanghai Institute of Plant Physiology and Ecology (SIPPE), Chines Academy of Sciences (CAS), the PPFD was fixed to 500 μmol m⁻² s⁻¹, and temperature 25 ± 1 °C, 70% RH and the photoperiod was 16 h light/8 h dark. The ABA was firstly dissolved in the distilled water to prepare a concentration of 40 μM ABA which was subsequently sprayed onto the mature leaves (2nd or 3rd pair of leaves counting from the top) of one-month old plants. During the ABA treatment, the mature leaves of control plants were sprayed only with distilled water (without ABA). For high light experiment, plants were grown under a phytotron in the Partner Institute of Computational Biology (PICB), CAS, with a PPFD of 1400 μmol m⁻² s⁻¹ (high light conditions) and that for the control condition was set to 500 μmol m⁻² s⁻¹. The high light condition was achieved by supplementing light with a lab-made light emitting diode (LED) light source. The light spectrum used in both control and high light conditions was as depicted in supplementary figure S6. For all experiments, plants were watered twice a week and fertilizer was weekly used after being dissolved to a concentration of 1% (w/w) (N: P: K = 20:20:20) to avoid plants growth limitation due to nutrition depletion and obtain healthy plants useful thereafter for our measurements.

PacBio Iso-seq and next-generation sequencing of RNA

To obtain a better annotation for Flaveria genes and understand their potential biological functions, we sequenced full-length transcript sequences across the transcriptome of *Fra* using Iso-seq, which was used as a reference to annotate genes of the other three species (supplementary fig. S1). The RNA samples isolated from four different tissues, including leaf, stem, root and flower, of *Fra* were mixed and subjected to the sequencing process in order to obtain a blueprint of transcripts. For leaf samples, the younger fully expanded leaf from one-month old plants, which usually located on the 2nd or the 3rd pair of leaves counting from the top, was used. For the stem samples, the mid-segment of a stem between the root node and that of the leaf from one-month old plants was used. Concerning the root samples, the root hair of one-month old plants was used. The compacted soil on the roots was removed by floating and gently shaking root in water before samples for RNA were taken. Regarding the flower samples, three to five inflorescences from the same plant were sampled. Each sample was collected from three different plants. All samples were cut and immediately frozen in liquid nitrogen. Total RNA for each tissue was then isolated following the protocol of the PureLInk™ RNA kit (ThermoFisher Scientific, USA). The isolated RNA from the above mentioned four
tissues was mixed in equal amount for the next step of analysis. The cDNA was synthesized using SMARTer® PCR cDNA Synthesis Kit (Clontech, USA) which was then amplified using the KAPA HiFi PCR Kits (Roche, USA). The amplified products were classified into different fragments according to their sizes (1~2 k, 2~3 k or > 3 k) using BluePippin Size Selection (Sage Science, USA) which were subsequently utilized to construct a library separately through the SMRTbellTM Template Prep Kit 1.0 (PacBio, USA). Libraries were sequenced by the PacBio Sequel (PacBio, USA).

For the NGS of the RNA data, the younger fully expanded leaf obtained from 4-week (ABA, high light and 2-week low CO₂), 6-week (4-week low CO₂) and 6.5-month (6-month low CO₂) old plants usually situated on the 2nd or 3rd pair of leaves counting started from the top was used for all the species studied herein. The chosen leaves were cut and immediately frozen into liquid nitrogen and stored thereafter at -80°C until further processing. The total RNA was isolated according to the procedure described above. The NGS of RNA data was performed in the Illumina platform in the paired-end mode with a read length of 150 bp.

Transcripts assembly, annotation and quantification

The Iso-seq long reads (LR) of Fra was corrected using proovread (Hackl, et al. 2014) with NGS of RNA-seq short reads (SR). In this process, the SR was mapped to LR and sequencing errors of LR were corrected by short-read-consensus approach (Hackl, et al. 2014). The rectified Iso-seq reads were then utilized to conduct transcript assembly applying IDP-de novo (Fu, et al. 2018). To define (or characterize) a reference on the gene scale, we used blast (V 2.2.31+) (Camacho, et al. 2009) to detect the transcript groups which have high similarity, i.e., those with an E-value threshold of 10⁻¹⁰ and the sequence identity threshold of 90%. The longest transcript in the transcript group was used as a reference gene. All the identified reference genes form together the gene reference dataset. The biological functions of Fra genes were annotated by seeking the orthologs in the database Uniprot (https://www.uniprot.org/). In this study, we used a combination of two datasets. The first represents the Uniprot reference clusters of Uniref50 (each entry represents a cluster of sequences with at least 50% sequence identity and 80% an overlap). For this first dataset, further details are available on the following website (https://www.uniprot.org/help/uniref). The second dataset constitutes the Swissprot (manually annotated and reviewed database). The Orthologs were predicted by using blast (V 2.2.31+) (Camacho, et al. 2009) with an E-value cutoff threshold of 10⁻⁵. The TFs were annotated by searching the orthologs of all plant TFs in the database PlantTFDB (http://plantfdb.cbi.pku.edu.cn/; (Riano-Pachon, et al. 2007) using blast (V 2.2.31+) (Camacho, et al. 2009) with an E-value threshold of 10⁻⁵ (supplementary fig. S1).

The transcript sequence of Fro, Fso and Fir were de novo assembled based on SR. The de novo assembly was also performed for the SR in Fra. A Trinity (version 2.8.4) (Grabherr, et al. 2011) was used to perform the de novo assembly using default parameters except constraining the transcript length to be no less than
300 bp. The *de novo* assembled transcripts were annotated by searching the orthologs from *Fra* genes using blast (V 2.2.31+) (Camacho, et al. 2009) with an E-value threshold of 10^-5.

The transcripts abundance was quantified as Transcripts per kilobase Per Million mapped reads (TPM) by applying the RSEM package (Li and Dewey 2011), where bowtie2 (version 2.3.4.3) (Langmead and Salzberg 2012) was used to map the SR to *de novo* assembled transcripts and the other parameters remained in default (supplementary figs. S2 and S3). The differentially expressed genes between, on one hand, the control and treated samples and, on the other hand, between the different species studied herein were calculated using edgeR package (Robinson, et al. 2010) with a *P*-value < 0.05 and FC > 1.5.

**qRT-PCR analysis**

To verify the gene expressions determined based on the SR *de novo* assembly and confirm the accuracy of our computational analysis, a qRT-PCR analysis was conducted on the same RNA samples used for NGS of RNA data from leaf of six-week old plants including four-week treatment to low CO\textsubscript{2} of 100 ppm. Between 0.2 and 0.5 μg RNA was used to reverse transcribe the first strand cDNA with Superscript II Reverse Transcriptase (TransGen Biotech, Beijing). The qRT-PCR mixture was prepared following the manufacturer’s instructions of UNICONTM qPCR SYBR Green Master Mix (YEASEN, Shanghai). Briefly, the mixture of cDNA, buffer and enzyme were pipetted into the Hard-Shell PCR 96-well plates (Bio-Rad, USA) and then covered by Microseal ‘B’ seal (Bio-Rad, USA). The qRT-PCR was run in the BIO-RAD CFX Connect-system (Bio-Rad, USA). Each PCR reaction volume of 20 μL contains 10 μL SYBR Green PCR Master Mix, 4 μL deionized H\textsubscript{2}O, 1 μL primers (0.5F + 0.5R) and 5 μL cDNA. The amplification reaction was initiated with a pre-denaturing step at 95 °C for 3 min followed by 39 cycles of denaturing at 95 °C for 10 s, annealing at 60 °C for 20 s, then extension at 72 °C for 30 s. The gene relative expression against the housekeeping gene, Actin7, was calculated as follows: 2^-ΔΔCT (ΔΔCT = CT, gene of interest-CT, Actin7), as described previously by (Livak and Schmittgen 2001). Data was processed using Bio-Rad CFX Maestro software (Bio-Rad, USA). For each gene, two technical replicates and three biological replicates were performed. The primers used for qRT-PCR analysis are listed in supplementary table S3.

**Isolation of 5’ flanking sequence from PEPC1 gene of four species**

The 5’ flanking sequence of the PEPC1 gene of the four Flaveria species (*Fro, Fso, Fra* and *Ftr*) was amplified from the total DNA through PCR reactions. The DNA was extracted following the protocol of TIANquick Midi Purification kit (TIANGEN Biotech, Beijing). The primers used for the DNA amplification procedure with PCR are listed in supplementary table S4. Ultimately, the PCR products were sequenced in collaboration with Sangon Biotech Company (Shanghai, China).

**A GRN construction**
The GRN was constructed with the CMIP software package (Zheng et al. 2016), which implements a path consistency algorithm based on the conditional mutual information (CMI) (Zhang, et al. 2012; Zheng, et al. 2016). The CMIP package uses a table of gene expressions as input. The method firstly computes zero-order network based on mutual information (MI), then eliminates the indirect relationship by considering the CMI, leading thereby to a first-order network. We also implemented the P-value determination in the CMIP package based on a permutation test. In other words, we shuffled 2000 times the expression level of each gene, generating thus 2000 null datasets. Thereafter, the P-value was defined as the proportion of CMI calculated from the null datasets greater than that calculated from the original dataset. We used a P-value of 0.001 as the cutoff during the genome-wide GRN reconstruction. The C4GRN was developed by only retaining the co-regulatory network of the C4 core metabolic genes and their positively co-regulated TFs (thereafter, C4TFs) from the genome-wide GRN.

In this study, special measures have been taken, thereafter, to ensure the development of a reliable and accurate GRN. In the original CMIP package, an adequate cutoff of gene-gene partial Pearson correlation coefficient (PCC) was automatically estimated by fitting the curve of interaction number in response to different cutoff based on an exponential function, and the cutoff was determined as the slopes intersection of the start and the end sections of the fitted curve (Zheng, et al. 2016). Given that we mainly focus on the C4GRN construction, we slightly modified this computational procedure. Specifically, we compared the relative abundance of all TF families (TF frequency) of the C4GRN under different PCC cutoffs, increasing from 0.1 to 0.9 with a step size of 0.1. Consequently, we found that the TF frequency maintains a high degree of similarity under different PCC cutoffs ranging from 0.1 to 0.7 (supplementary fig. S7). We calculated the similarity of TFs frequencies under two adjacent PCC cutoffs using Pearson correlation, and found the Pearson correlations are high (>0.95) from 0.1 to 0.7 but plummets from 0.7 to 0.8 (supplementary fig. S8). We thus selected 0.7 to be the PCC cutoff for the GRN construction. Interestingly, the number 0.7 represents the inflexion point on the curve, which (the curve) depicts the edge number versus different PCC cutoffs (supplementary fig. S9).

Genome-wide GRN clustering and gene ontology enrichment analysis

In order to assess the potential implications of C4 genes and C4TFs in the different species such as their biological functions and/or the molecular relevance of their concurrent co-regulations, the genome-wide GRN was clustered into different modules. The latters (modules) enriched in C4TFs or those including the maximum C4 genes were selected for the functional enrichment analysis. Herein, we used a Markov Cluster Algorithm (MCL) package (version 14-137; (Dongen 2008) to detect the modules in the GRN. The MCL is an unsupervised model, where the number of clusters cannot be determined on the fly. The parameter “I” (inflation) represents the main handle affecting the cluster granularity and influences the cluster number and clustering efficiency that the algorithm produces (Enright, et al. 2002). The clustering efficiency was automatically calculated in the program by considering two factors. These are; the density of all clusters and the percentage of lost interactions. In order to determine a proper value for the “I” parameter used in MCL...
package, we examined the clustering efficiencies and cluster numbers for five different cutoffs of “I” including 1.4, 2, 4, 5, and 6. Then, we have chosen the ideal value for “I” which had the maximum clustering efficiency for a minimum clusters number. We found that all the four species investigated herein show the best “I” to be 2 (supplementary fig. S10A), with a clustering efficiency being 0.21, 0.2, 0.16 and 0.22 for *Fro, Fso, Fra* and *Ftr*, respectively.

Subsequently, we determined the modules enriched in the C₄ TFs (MET) using the Fisher’s test computed in-home R script. The calculated *P*-values from the Fisher’s test were adjusted using the Benjamini-Hochberg (BH) method (Benjamini and Hochberg 1995) integrated in R software (version 3.0.2). A threshold of 0.001 for the adjusted *P*-values was used. Besides, the modules including the maximum of C₄ orthologous gene were termed as MMC₄. To investigate the potential involvements of MET and MMC₄ in certain biological functions, we calculated the over-represented gene ontology (GO) for each MET and MMC₄ using Fisher’s test with BH adjusted *P* < 0.001, where the genes in the genome-wide GRN were used as a background.

**Accession number**

The RNA-seq and Iso-seq data were submitted to Gene Expression Omnibus (GEO) in the Nation Center for Biotechnology Information (NCBI) database under the following accession number: GSE143470.

**Author contributions**

XGZ, GC, MJL and JE designed the project and wrote the paper, MJL and FC did bioinformatics analysis, MJL did the qRT-PCR and JE did the RNA-isolation.

**Competing interests**

The authors have no competing of interests to declare.

**Acknowledgements**

The authors thank and appreciate Prof. Rowan F. Sage and Peter Westhoff for sharing with us Flaveria materials. We thank Prof. Haiyang Hu and Mingnan Qu for their fruitful discussions and suggestions. Authors thank also Fengfeng Miao, Qingfeng Song and Xinyu Liu for their assistance in maintaining the growth chambers conditions properly and according to the recorded instructions. This work was supported by National Science Foundation of China, NSFC (31701139, 31870214, 31500988), Strategic Priority Research Program of the Chinese Academy of Sciences, CAS (Grant No. XDB27020105), National Research and Development Program of Ministry of Science and Technology of China, MSTC (2019YFA0904600).

**Supplementary Material**

SFI: supplementary figures and tables
SF2: Functional annotations of *Fra* genes

SF3: RNA-seq mapping statistics

SF4: Gene expression in TPM

SF5: C₄ GRN tables.

SF6: C₄ TFs enriched GOs

SF7: PEPC1 promoter sequences of the four Flaveria species

SF8: Functional annotations of the 37 C₄ TFs overlapped between *Fso* and *Ftr*
References


A mem1


Figure 1. Sample information and flowchart of constructing gene co-regulatory network

(A) A schematic representation of the evolutionary relationships between four Flaveria species and their respective photosynthetic types studied herein. These are C₃, type I C₃-C₄ (I C₃-C₄), type II C₃-C₄ (II C₃-C₄) and C₄. (B) Flowchart of constructing gene co-regulatory network (GRN) based on RNA-seq data. In addition to RNA-seq data, we also generated a paralleled Iso-seq for Fra to improve gene annotation for Flaveria species. Thus, we constructed a genome-wide GRN using the conditional mutual information (CMI), and then we extracted the sub-network which contains C₄ genes and their positively co-regulatory TFs, termed thereafter as C₄GRN. (C-D) Principle plots of all the RNA-seq data set based on gene expression profiling. Fra and Fro clustered together when all species were plotted simultaneously (C); however, they are separated when solely their RNA-seq data were plotted (D). Abbreviations: Fro: F. robusta, Fso: F. sonorensis, Fra: F. ramosissima, Ftr: F. trinervia, GRN: gene co-regulatory network.
Figure 2. C₄ core metabolic pathway and its involved genes

(A) The histogram plots depict the quantification of the determined C₄ version of C₄ genes based on qRT-PCR analysis performed on leaf samples of plants grown at low CO₂ concentration (100 ppm) for four weeks. (B) The diagram of the core C₄ pathway and the involved genes in Flaveria. Ftr is a typical NADP-ME type of C₄ photosynthesis. The heatmaps show the log₂-transformed fold changes of transcript levels of C₄ genes assessed as Transcript per kilobase Per Million mapped reads (TPM) in each species compared with their counterparts in the C₃ species Fro. RNA-seq quantification from 4-week low CO₂ experiment is showed here. A star (*) represents significant changes at a P-value < 0.05, and FC > 1.5. Abbreviations: CA1, carbonic anhydrase 1; PEPC1, phosphoenolpyruvate carboxylase 1; PEPC-K, PEPC kinase; NADP-MDH, NADP-dependent malate dehydrogenase; NADP-ME4, NADP-dependent malic enzyme 4; AlaAT2, alanine aminotransferase 2; PPDK, pyruvate/orthophosphate dikinase; PPDK-RP: PPDK regulatory protein. Metabolite transporters were displayed in the encircled numbers located on the membrane, with the consecutive number from 1 to 4 refer to DiT1 (dicarboxylate transporter 1), DiT2 (dicarboxylate transporter 2), BASS2 (bile acid sodium symporter 2), NHD1 (sodium: hydrogen antiporter 1) and PPT1 (phosphate/phosphoenolpyruvate translocator 1), respectively. Species abbreviations are as depicted in Fig. 1.
Figure 3. The C₄GRN properties

(A) Basic information on the genome-wide GRN and C₄GRN in four Flaveria species. (B) The overlapped TFs in the C₄GRN between the four investigated Flaveria species. (C) The C₄GRN in the C₄ species Ftr was displayed, where only the interactions between the C₄ genes and their co-expressed TFs are shown. The C₄ genes were depicted in red font and the TFs were shown in circles with 4 different colors. The overlapped TFs with other three species were exhibited in different colors as depicted by the legend “TF” below the panel. The overlapped C₄ genes-TFs interactions were displayed in the same colors used for the overlapped TFs as reflected by “edge” in the legend. (D) The expression patterns of the 196 co-regulated TFs with the C₄ genes in Ftr were exhibited in a heatmap, which was calculated based on the RNA-seq data obtained from plants, of each species, experienced low CO₂ (100 ppm) for four weeks. Colors represent the relative transcript abundance, with red, white and black mean high, median and low expression level, respectively. The comparison of the TFs with the highest transcript abundance in Fso to those of the three other species was delimited by the selected zone with the blue dashed-edge rectangle on the heatmap. (E) Word clouds showing the C₄TFs frequency in different TF
families, with large word size indicates the higher frequency. (F) Co-regulation network of C4 genes and TFs from GRAS family in Ftr. The colors of circles and edges are same as in the legend of panel (C). Species abbreviations are the same as Fig. 1.
Figure 4. Biological functions of genes from MET and MMC4 in different species

(A) A predictive diagram displaying the analysis pipeline used to identify the potential functions of C4 TFs. A genome-wide GRN was firstly clustered into modules. The latter (modules) enriched in C4 TFs (Fisher’s test, \( P < 0.001 \), BH adjusted) were then determined and termed modules enriched in C4 TFs (MET). Over-represented GO functions of genes in the MET were calculated using Fisher’s test (\( P < 0.001 \), BH adjusted).

(B) Bar plots show the number of C4 TFs and C4 genes in the modules that encompass C4 genes. Bars in antique white represent the number of C4 TFs and bars in orchid represent the number of C4 genes. The number given in grey and black show the number of C4 TFs (upper value) and total number of genes (bottom value) in each module, and the numbers in orchid show the number of C4 genes. The star “*” indicates that the module was a MET. Over-represented GO terms of genes from MET (C) and genes from modules encompassing the maximum C4 genes (MMC4) (D) were shown.

Abbreviations: BP, molecular process; CC, cell component; MF, molecular function. The species abbreviations are as depicted in Fig. 1. BH means Benjamini and Hochberg.
Figure 5. The C₄GRN overlap between the intermediate and C₄ species

(A) and (B) show the sub-networks of C₄GRN in Fso and Fra, respectively. These show the regulatory relationship between C₄ gene and the shared C₄ TFs with Ftr. The C₄ genes are depicted in red font. TFs are shown in circles, with red, blue and grey colors represent the TFs showing an increased, decreased, or similar (comparable) transcript abundance in the C₄ species Ftr compared to that in the intermediate species (edgeR, P < 0.05 and FC > 1.5). Four TFs co-regulating the PEPC1 are found to be conserved between Fra and Ftr, as labeled by TF1-4, which is from TF family of MYB_related, NF-YA, GRAS and DBB, respectively. (C) The mesophyll expression module 1 (MEM1) sequences of the PEPC1 promoters from four species. The A and B submodules are highlighted in boxes. Asterisks show identical nucleotides in the two modules. Red zones indicate the single nucleotide difference in the A submodule, and the required tetranucleotide CACT in the B submodule. The full promoter sequences are available in Supplementary Material SF7. Abbreviations of enzymes are same as those given in Fig. 2.
Figure 6. The C₄GRN evolution towards a C₄ photosynthesis stage

(A) In Ftr, 93 C₄GRN TFs show higher transcript abundance than Frob (edgeR, P < 0.05 and FC > 0.5), termed as C₄-required TFs (C₄ReTFs), which co-regulate all the 13 C₄ core metabolic genes. (B) 37 of C₄ReTFs are present in the C₄GRN of Fso, co-regulated with all C₄ core metabolic genes except PEPC1. TFs are shown in circles and triangles, with triangles presenting TFs shared with either Fso (panel A) or Fro (panel B). Red, blue, and grey colors represent TF showing higher, lower and similar transcript abundance compared to Fso (A) or Fro (B) (edgeR, P < 0.05 and FC > 1.5). (C) Seven among the 37 TFs present in the C₄GRN of Fro which (seven TFs) co-regulate three orthologs of C₄.
The CI-GRN of Ftr and Fso are shown in Fig. S5. Abbreviations of species and enzymes are same as those given in Fig. 2.
Figure 7. A proposed model for the C₄GRN evolution in Flaveria genus

Two key stages are involved during the C₄GRN evolution in Flaveria genus. Though the C₄ genes orthologs were co-regulated with the photosynthesis related genes in C₃ species, the TFs required for the C₄ photosynthesis remain still absent. During the transition from the C₃ species to type I C₃-C₄ species, a large number of C₄-required TFs (C₄ReTFs) were recruited, but not for all C₄ genes. The second stage constitutes the transition from type I C₃-C₄ species to C₄ species, where a complete C₄GRN establishment could be achieved. This C₄GRN includes the recruitment of new TFs, enhancement of the transcript abundance of C₄ReTFs that are already recruited in type I C₃-C₄ species and the formation of new regulations between the already pre-existing TFs and other C₄ genes. Fra may represent a divergent CO₂ CCM as C₄ species, in which the C₄GRN is different from those of type I C₃-C₄ and C₄ species. PEPC and PPDK-RP show a greater number of co-regulated TFs than other C₄ core metabolic genes in Ftr.
### Table 1.
RNA-seq data generated for constructing GRN; hrs, wks and mths stand for hours, weeks and months, respectively.

<table>
<thead>
<tr>
<th>Species</th>
<th>low CO₂ 2 wks</th>
<th>low CO₂ 4 wks</th>
<th>low CO₂ 6 mths</th>
<th>ABA 3 hrs</th>
<th>high light 2 wks</th>
<th># sample</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Fro</em></td>
<td>6</td>
<td>6</td>
<td>4</td>
<td>6</td>
<td>6</td>
<td>28</td>
</tr>
<tr>
<td><em>Fso</em></td>
<td>6</td>
<td>6</td>
<td>4</td>
<td>6</td>
<td>n.a</td>
<td>22</td>
</tr>
<tr>
<td><em>Fra</em></td>
<td>6</td>
<td>6</td>
<td>n.a</td>
<td>6</td>
<td>6</td>
<td>24</td>
</tr>
<tr>
<td><em>Ftr</em></td>
<td>6</td>
<td>6</td>
<td>n.a</td>
<td>6</td>
<td>6</td>
<td>24</td>
</tr>
</tbody>
</table>