Whole-Genome Duplication Facilitated the Evolution of C4 Photosynthesis in Gynandropsis gynandra

Chi-Fa Huang,1 Wen-Yu Liu,1 Mei-Yeh Jade Lu,1 Yi-Hua Chen,1 Maurice S.B. Ku,2 and Wen-Hsiung Li*,1,3

1Biodiversity Research Center, Academia Sinica, Taipei, Taiwan
2Department of Bioagricultural Science, National Chiayi University, Chiayi, Taiwan
3Department of Ecology and Evolution, University of Chicago, Chicago, IL, USA

*Corresponding author: E-mail: whli@uchicago.edu.
Associate editor: Michael Purugganan

Abstract

In higher plants, whole-genome duplication (WGD) is thought to facilitate the evolution of C4 photosynthesis from C3 photosynthesis. To understand this issue, we used new and existing leaf-development transcriptomes to construct two coding sequence databases for C4 Gynandropsis gynandra and C3 Tarenaya hassleriana, which shared a WGD before their divergence. We compared duplicated genes in the two species and found that the WGD contributed to four aspects of the evolution of C4 photosynthesis in G. gynandra. First, G. gynandra has retained the duplicates of ALAAT (alanine aminotransferase) and GOGAT (glutamine oxoglutarate aminotransferase) for nitrogen recycling to establish a photorespiratory CO2 pump in bundle sheath (BS) cells for increasing photosynthesis efficiency, suggesting that G. gynandra experienced a C3–C4 intermediate stage during the C4 evolution. Second, G. gynandra has retained almost all known vein-development-related paralogous genes derived from the WGD event, likely contributing to the high vein complexity of G. gynandra. Third, the WGD facilitated the evolution of C4 enzyme genes and their recruitment into the C4 pathway. Fourth, several genes encoding photosystem I proteins were derived from the WGD and are upregulated in G. gynandra, likely enabling the NADH dehydrogenase-like complex to produce extra ATPs for the C4 CO2 concentration mechanism. Thus, the WGD apparently played an enabler role in the evolution of C4 photosynthesis in G. gynandra. Importantly, an ALAAT duplicate became highly expressed in BS cells in G. gynandra, facilitating nitrogen recycling and transition to the C4 cycle. This study revealed how WGD may facilitate C4 photosynthesis evolution.

Key words: C4 photosynthesis, C4 evolution, whole-genome duplication, comparative genomics.

Introduction

Kranz anatomy is a distinctive structure of C4 leaves in which the vein is wrapped around by one inner layer of bundle sheath (BS) cells and then one outer layer of mesophyll (M) cells (Hatch 1987). This structure has allowed the evolution of a CO2 concentration mechanism (CCM) that transports CO2 from M to BS cells for its final fixation through the Calvin cycle in BS cells (Hatch 1987). Compared with C3 plants, this unique structure coupled with a high vein density in leaves confers C4 plants a superior photosynthesis efficiency with increased tolerance to high light, heat, and drought. The Kranz anatomy with the special CCM is a well-known example of convergent evolution in plant evolution, which has occurred in over 60 plant lineages (Sage et al. 2011).

Increased leaf vein density and development of Kranz leaf anatomy have been considered an early primary step in C4 evolution (Sinha and Kellogg 1996; Gowik and Westhoff 2011; Christin et al. 2013). From a comparative study of the transcriptomes of developing leaves in C3 Tarenaya hassleriana and C4 Gynandropsis gynandra, we proposed that elevated auxin biosynthesis and transport are responsible for the development of high vein density in C4 leaves (Huang et al. 2017).

Various models, including anatomical, physiological, phylogenetic, and computational modeling, predicted that confinement of photorespiration in BS cells, to raise CO2 concentration at the site of Rubisco (ribulose-1,5-bisphosphate carboxylase/oxygenase) in BS cells, provides an important intermediate step in C3 to C4 evolution (Monson 1999; Bauwe 2011; Heckmann et al. 2013; Williams et al. 2013; Mallmann et al. 2014). Rubisco is the key enzyme of photosynthetic carboxylation reaction via the Calvin–Benson or C3 cycle. However, photorespiration, also known as the oxidative photosynthetic carbon cycle, may occur simultaneously with carboxylation on Rubisco, leading to CO2 release in mitochondria and thus causing loss of fixed carbon (Sharkey 1988). The key enzyme for releasing fixed CO2 in mitochondria is glycine decarboxylase (GDC) (see supplementary table S1, Supplementary Material online for gene name abbreviations and functions) comprising the P-, L-, T-, and H-proteins in mitochondria. The GDC P-protein (GLDP) is the decarboxylase that catalyzes the decarboxylation of glycine to release...
CO₂ (Oliver and Raman 1995). In C₃ species, CO₂ released from photorespiration in M cells diffuses out of the leaf, thus reducing net photosynthesis with a high CO₂ compensation point. In C₃–C₄ intermediate species, glycine derived from photorespiration serves as a CO₂ carrier and is shuttled from M to BS cells where GLDP decarboxylates glycine to release CO₂ for refixation, which is called the photorespiratory CO₂ pump in C₃–C₄ plants (fig. 1) (Hylton et al. 1988). Due to the elevated CO₂ concentration in BS cells, carboxylation by BS Rubisco is favored over oxidation, leading to reduced photorespiratory CO₂ loss and an overall increase of photosynthesis efficiency. At the same time, a basic C₄ cycle is recruited to prevent nitrogen imbalance created by the photorespiratory CO₂ pump in C₃–C₄ plants (fig. 1) (Mallmann et al. 2014).

At the final step of C₄ evolution, C₄ enzyme genes are upregulated and recruited to participate in the operation of C₄ cycle between well differentiated M and BS cells (fig. 1). After the establishment of the C₄ cycle, the C₄-type CCM replaces the photorespiratory CO₂ pump to concentrate CO₂ in the BS cells of C₄ plants (fig. 1).

Compared with C₃ photosynthesis, C₄ photosynthesis requires two extra ATPs to drive the CCM for each CO₂ molecule fixed, and the cyclic electron flow (CEF) around photosystem I (PSI) was predicted to contribute the additional ATPs required (Munekage et al. 2004). Two distinct CEF pathways, NADH dehydrogenase-like (NDH) complex- and ferredoxin: plastoquinone oxidoreductase (FQR)-dependent flows, have been identified in C₃ plants (Munekage et al. 2002; Ifuku et al. 2011). Both pathways transfer excited electrons to the cytochrome b₆f complex, which pumps protons into the thylakoid space (Wikstrom et al. 1981). These pumped protons contribute to the electrochemical proton gradient across the thylakoid membrane of chloroplasts, which is then used to drive ATP synthesis.

Gene duplication, either single or whole-genome duplication (WGD), has been proposed to be a prerequisite for C₄ evolution (Monson 2003) because it provides extra gene copies to reduce selective constraint and to acquire beneficial morphological or biochemical modifications (Panchy et al. 2016). A previous study showed that a photorespiratory GLDP experienced duplication in ancestral C₃ Flaveria species. One copy of the GLDP duplicates became preferentially expressed in BS cells and helped to establish the photorespiratory CO₂ pump in Flaveria C₃–C₄ intermediate species (Schulze et al. 2013). In addition, several C₃ enzyme genes, including PEPC, PPDK, NADP-ME, NADP-MDH, and CA, have undergone duplication in the ancestral C₃ Flaveria plants. One copy of each C₃ gene duplicate pair was subsequently upregulated and modified for organelle-, cell-, and organ-specific expression to support the CCM of the NADP-ME subtype C₃ photosynthesis in Flaveria (Monson 2003).

Cleomaceae, a sister family of Brassicaceae, has undergone the evolution of C₄ photosynthesis at least three times (Bayat et al. 2018). Cleomaceae and Brassicaceae shared two common ancient WGD events, namely the β and γ WGD events (Barker et al. 2009). In addition, both lineages underwent an independent WGD event after their divergence, called Th-α in the Cleomaceae lineage and At-α in the Arabidopsis lineage (Barker et al. 2009). The At-α and Th-α WGD events were estimated to occur ~34 and ~20 Ma (million years ago), respectively (Bayat et al. 2018), whereas At-β occurred ~170–235 Ma (Bowers et al. 2003). Because C₄ photosynthesis in G. gynandra evolved after the Th-α WGD event (van den Bergh et al. 2014), it provides a good model for studying the role of a WGD event in the evolution of C₄ photosynthesis. Also, it may help explore why C₄ photosynthesis did not evolve in T. hasslerianna, although it shared the Th-α WGD event.

In this study, we first used PacBio Iso-Seq to obtain long transcripts of G. gynandra genes to assemble a CDS (coding sequence) data set and we then added the de novo CDSs assembled from the illumina RNA-seq reads (Huang et al. 2017) to complement the PacBio Iso-Seq data. We also constructed a T. hasslerianna CDS data set from the previously predicted CDSs of T. hasslerianna (Cheng et al. 2013) and the de novo CDSs assembled from Illumina RNA-seq reads (Huang et al. 2017). After constructing two CDS databases for the two species, we used the RNA-seq reads of six leaf-development stages (S0–S5) deposited at NCBI (Kulahoglu et al. 2014) to calculate expression levels of the CDSs. S0 was the stage with leaves ~2 mm in length. Then, every two consecutive stages were separated by 2 days. The leaves initiated secondary vein formation at S1 and fully developed by S4 and S5. The data were then used to identify paralogs, date the duplication time, and calculate the ratio of the nonsynonymous substitution rate to the synonymous substitution rate (i.e., the K_a/K_s ratio). Using these data, we explored the relationship between the Th-α WGD event and the evolution of C₄ photosynthesis in G. gynandra. This study revealed that the Th-α WGD event played a crucial role in the evolution of C₄ photosynthesis in G. gynandra, including anatomical and biochemical modifications. Also, this study provides insights into why C₄ photosynthesis failed to evolve in T. hasslerianna, although it shared the Th-α event with G. gynandra.

### Results

#### cDNA Assembly

The PacBio sequencing of the G. gynandra CDNA library produced 662,111 raw reads, with an average length of 1,854 bp (supplementary table S2, Supplementary Material online). Our CDS assembly procedure is illustrated in supplementary figure S1, Supplementary Material online, and explained in detail in Materials and Methods. After the pipeline processing involving circular consensus sequencing (CCS), read classification, transcript clustering and polishing using isoSeq3, we obtained 39,766 polished high-quality (HQ) isoforms (supplementary table S2, Supplementary Material online). Then, we combined the PacBio and Illumina RNA-seq transcript data set to obtain a G. gynandra CDS data set containing 21,345 ORFs, including 15,431 full-length CDSs (supplementary data set S1a, Supplementary Material online). The number of CDS sequences inferred from PacBio reads alone was 1,454, that from Illumina reads alone was 9,325 and that from both PacBio and Illumina reads was 10,566; the total number
was 21,345. In *T. hassleriana* only Illumina reads were available, and we obtained a CDS database that contains 27,500 CDSs including 21,162 full-length CDSs (supplementary data set S1b, Supplementary Material online). Fewer CDSs and full-length CDSs were inferred in *G. gynandra* than in *T. hassleriana*, because the genome of *G. gynandra* has not yet been sequenced, making it more difficult to infer coding sequences in *G. gynandra*.
Identification of Duplicate Genes Derived from the Th-α WGD Event

To identify the duplicate (paralogous) genes in a species, we used reciprocal BLASTp searches with \(E \leq 1.0 \times 10^{-5}\). In *T. hassleriana* and *G. gynandra*, the set of duplicate genes thus obtained includes duplicate genes derived from not only the Th-α WGD event but also the Th-β WGD event. These two groups of duplicate genes may be separated by using the \(K_S\) values between homologous genes, where \(K_S\) is the number of synonymous substitutions per synonymous site. Barker et al. (2009) estimated that in *T. hassleriana*, the median \(K_S\) value for duplicate genes derived from the Th-α WGD (denoted as “Th-α median \(K_S^{T}\)” is 0.41 (0 < \(K_S < 1.1\)) and Th-β median \(K_S = 1.68\) (1.1 ≤ \(K_S < 2.1\)). Thus, we used \(K_S = 1.1\) to divide the paralogs of *T. hassleriana* (*G. gynandra*) into a group of Th-α paralogs with \(K_S < 1.1\) and a group of Th-β paralogs with \(1.1 \leq K_S < 2.1\). The criterion of \(K_S < 1.1\) assumes that there is no Th-α paralog pair in *T. hassleriana* and *G. gynandra* that has \(K_S > 1.1\). It will be seen that this assumption holds well when we consider the distribution of \(K_S\) values in the next subsection. By this separation criterion, we obtained for *T. hassleriana* 6,787 duplicate-gene pairs derived from the Th-α WGD and 295 pairs derived from the Th-β WGD (supplementary data set S2, Supplementary Material online). The corresponding values for *G. gynandra* are 3,454 and 224 pairs (supplementary data set S2, Supplementary Material online). Note that in both species, the Th-α duplicate-gene pairs also include duplicate-gene pairs derived from non-WGD duplications. This issue will be discussed in the next subsection.

Distribution of \(K_S\) Values between Paralogous Genes in a Genome

The \(K_S\) value between two homologous genes is usually not strongly affected by natural selection and thus can give a sense of the divergence time between the two genes. We therefore computed the \(K_S\) values between paralogous genes in *G. gynandra* and in *T. hassleriana*. For this purpose, we used the KaKsAnalysis tool of PlantTribes 2 (produced by the AssemblyPostProcessor) in PAML (Yang 2007), which gives not only \(K_S\) but also \(K_A\) (number of nonsynonymous substitutions per nonsynonymous site). As our CDGs for *G. gynandra* were constructed solely from leaf transcriptomes, for both *G. gynandra* and *T. hassleriana*, we used only genes that are expressed in leaves. Here, a gene is said to be expressed if its RPKM (reads per kilobase per million mapped reads) is ≥ 1 in the RNA-seq data (Külahoglu et al. 2014); we identified 3,745 and 7,197 sets of expressed paralogous pairs in *G. gynandra* and *T. hassleriana*, respectively (supplementary data sets S1 and S2, Supplementary Material online). Figure 2 shows the distribution of \(K_S\) values for the paralog pairs in *T. hassleriana* and also the distribution in *G. gynandra*; the \(K_S\) values are given in supplementary data set S2, Supplementary Material online. Note that the two distributions are very similar to each other with a slight shift of the former distribution to the left; a two-sample Kolmogorov–Smirnov test found no significant difference between the two \(K_S\) distributions (\(P \text{ value} = 0.28\)). Note further that the two distribution curves decrease to almost 0 when \(K_S = 1.025\). Thus, our assumption that there is no Th-α paralog pair with \(K_S > 1.1\) in *G. gynandra* and *T. hassleriana* holds approximately. From the two distributions we obtained Th-α median \(K_S = 0.41\) and Gg-α median \(K_S = 0.48\), which agree well with the estimate of Th-α median \(K_S = 0.41\) in *T. hassleriana* by Barker et al. (2009).

We also computed the \(K_S\) values of orthologous genes between *T. hassleriana* and *G. gynandra*. For simplicity, we considered only single-copy orthologs in the two species (supplementary data set S2, Supplementary Material online), which were identified using the OrthoFinder tool (Emms and Kelly 2019). Figure 2 shows that the distribution of \(K_S\) values between *G. gynandra* and *T. hassleriana* single-copy genes is very close to the distributions of \(K_S\) values for the Th-α paralog pairs in *G. gynandra* and *T. hassleriana*. This implies that the speciation between *G. gynandra* and *T. hassleriana* occurred soon after the Th-α WGD event.

Figure 2 also includes the distributions of \(K_S\) values between paralogs in *T. hassleriana* and in *G. gynandra* derived from the Th-β WGD event. We estimated Th-β median \(K_S = 1.54\) and Gg-β median \(K_S = 1.48\), which are somewhat smaller than Barker et al.’s estimate of Th-β median \(K_S = 1.68\).

The groups of Th-α and Gg-α paralogs likely included a substantial number of non-WGD duplicates as indicated by a bump in the distribution curve in the end region with \(K_S < 0.2\) (fig. 2). In *Arabidopsis thaliana*, we identified 3,009 duplicate pairs with \(K_S < 1.1\), using the KaKsAnalysis tool of PlantTribes 2 in PAML (Yang 2007). Among them we found 896 pairs in the set of non-WGD (tandem) duplicate pairs and the remaining 2,113 pairs in the set of At-α WGD duplicate pairs in Wang et al. (2013). A bump in the end region with \(K_S < 0.55\) is seen in the distribution of \(K_S\) values for the pool of non-WGD and Th-α WGD duplicate pairs (3,009 pairs), when compared with the distribution of \(K_S\) values for the At-α WGD duplicate pairs (2,113 pairs) alone (supplementary fig. S2, Supplementary Material online). As mentioned above, the Th-α WGD and the At-α WGD were estimated to occur 20 and 34 Ma, respectively. If we assume that non-WGD occurs at a constant rate in *T. hassleriana*, then the number of non-WGD duplicates that occurred after the Th-α WGD is estimated to be \((20/34) \times 896 = 527\). In *T. hassleriana*, the pool of non-WGD and Th-α WGD duplicate pairs is found to be 6,787. Thus, the proportion of non-WGD duplicate pairs in the pool of non-WGD and Th-α WGD duplicate pairs (527/6,787 = 0.08) in *T. hassleriana* is considerably smaller than that in *A. thaliana* (896/3,009 = 0.30). For this reason, in *T. hassleriana* the contribution of non-WGD duplicate pairs to the pool of non-WGD and Th-α WGD duplicate pairs is likely less than 10%. This may explain why the distribution curve denoted by Th-α in figure 2, which includes both non-WGD and Th-α WGD duplicate pairs, is much sharper than that for At-α, although the latter includes only At-α WGD duplicate pairs. However, in principle, in the absence of recent non-WGD duplicates, the two distribution curves for Th-α and Gg-α should decrease to almost 0 when
KS values between paralogs or between orthologs. Th-α (Gg-α) refers to the distribution of KS values between paralogs in *Tarenaya hassleriana* (*Gynandropsis gynandra*) derived from the Th-α WGD event; the paralogs actually also include non-WGD duplicates. At-α refers to the distribution of KS values between paralogs in *Arabidopsis thaliana* derived from the At-α WGD event. "Gg versus Th" refers to the distribution of KS values between G. gynandra and T. hassleriana single copy genes. "At versus Th" refers to the distribution of KS values between A. thaliana and T. hassleriana single copy genes. Th-β (Gg-β) refers to the distribution of KS values between paralogs in T. hassleriana (*G. gynandra*) derived from the Th-β WGD event. The number (n) of gene pairs used: Gg-α: n = 3,454 paralogous pairs; Th-α: n = 6,787 paralogous pairs; Gg versus Th: n = 5,840 orthologous pairs of single-copy genes; At-α: n = 2,113 paralogous pairs; Gg-β: n = 224 paralogous pairs; Th-β: n = 295 paralogous pairs; At versus Th: n = 3,064 orthologous pairs of single copy genes.

KS values become smaller than 0.2 as in the distribution of KS values between *G. gynandra* and *T. hassleriana* single-copy genes (fig. 2). Therefore, we may assume that most or nearly all paralogs with KS < 0.2 in *G. gynandra* and *T. hassleriana* were derived from non-WGD duplications.

Figure 2 also includes the distribution of KS values between duplicate genes in *A. thaliana* derived from the At-α WGD event. This distribution is some distance from the left of the distribution of KS values for the paralog pairs in *G. gynandra*, suggesting that the At-α WGD occurred considerably earlier than the Th-α WGD. It is, however, close to the distribution of KS values between *A. thaliana* and *T. hassleriana* single-copy genes, suggesting that the At-α WGD event occurred soon after the divergence between the *A. thaliana* and the *T. hassleriana* lineage.

Duplicates of C₄-Related Genes in *G. gynandra* and *T. hassleriana*

In *Flaveria* (yellowtops), several C₄ enzyme genes, including PEPC, PPDK, NADP-ME, NADP-MDH, and CA, have undergone duplication in the ancestral C₃ plants and subsequent neo-functionalization in C₄ photosynthesis of *Flaveria* species (NADP-ME subtype) with higher expression levels (Monson 2003). In *G. gynandra*, a C₄ plant belonging to the NAD-ME subtype C₄ plants, we found eight genes (βCA2, βCA4, PEPC2, NAD-ME2, ALAAT, MMDH, TPT, and PPT) in the C₄ photosynthesis pathway that have at least two copies in *G. gynandra* (fig. 3); the first six genes encode enzymes whereas the last two encode triophosphate and PEP transporters. For all eight genes, the paralog pairs in both *G. gynandra* and *T. hassleriana* have KS values smaller than 0.85 (supplementary table S3, Supplementary Material online), which is considerably smaller than the median value (1.48) for the paralog pairs derived from the Th-β WGD (fig. 2). Thus, none of these paralog pairs was derived from the Th-β WGD. Below, we investigate whether these paralogs were derived from the Th-α WGD.

Figure 3 shows the neighbor-joining (NJ) trees based on nonsynonymous substitutions for the eight genes in *G. gynandra* and *T. hassleriana*, with the homologous genes in *A. thaliana* as the outgroup. Figure 3A shows the NJ tree for the βCA2 genes. Note that GgβCA2a and GgβCA2b are closely related with a KS value of only 0.10 (supplementary table S3, Supplementary Material online), which is much smaller than 0.48, the Gg-α median KS, suggesting that these two genes were derived from a recent non-WGD duplication. Their common ancestor and ThβCA2b form an ortholog pair supported by a bootstrap value of 86% (fig. 3A). Moreover, GgβCA2 and ThβCA2a form another ortholog pair supported by a bootstrap value of 98%. Thus, GgβCA2 and the common ancestor of GgβCA2a and GgβCA2b were apparently derived from the Th-α WGD and so were ThβCA2a and ThβCA2b. This conclusion is supported by the maximum-likelihood (ML) trees based on nucleotide and amino acid sequences (supplementary fig. S3, Supplementary Material online). The KS value between GgβCA2a (GgβCA2b) and GgβCA2 is 0.57 (0.58), which is larger than the Gg-α median KS (0.48). The KS value between ThβCA2a and ThβCA2b is 0.42, which is almost the same as the Th-α median KS (0.41). Thus, the KS values support the view that the above paralog pairs were derived from the Th-α WGD.

Figure 3E shows the NJ tree for the ALAAT genes. It is identical in topology to both of the ML trees in supplementary fig. S3, Supplementary Material online, though it is different from the NJ tree based on synonymous substitutions. The single ALAAT gene in *T. hassleriana* is clustered with GgALAAT in three of the four trees (fig. 2 and supplementary fig. S3, Supplementary Material online), so it is likely the ortholog of GgALAAT. The KS value between GgALAAT and GgALAAT is 0.68, which is larger than the Th-α median of 0.41, so it is reasonable to assume that they were derived.
from the Th-x WGD rather than from a non-WGD duplication.

**Figure 3B** shows the NJ tree for the CA4 genes. The phylogenetic positions of the three CA4 genes in *T. hassleriana* differ among the four trees in **Figure 3B** and supplementary figure S3, Supplementary Material online. However, as the $K_{A}$ (0.05) and $K_{S}$ (0.45) values between ThβCA4a and ThβCA4b are smaller than those between ThβCA4a and ThβCA4c (0.11 and 0.55) and between ThβCA4b and ThβCA4c (0.09 and 0.60), we may assume that ThβCA4a and ThβCA4b were derived from a more recent duplication than the duplication that produced their common ancestor and ThβCA4c, which...
was likely derived from the Th-α WGD because the Ks values between ThβCA4a and ThβCA4b (0.55 and 0.60) are larger than the Th-α median (0.41). Similarly, as the Ks value between GgβCA4a and GgCAβCA4 (0.67) is larger than the Gg-α median (0.48), we assume that the two genes were derived from the Th-α WGD.

Like the case of CA4 genes, although the phylogenetic trees for the other five genes in figure 3 give no clear evidence that the two paralogs in G. gynandra were derived from the Th-α WGD, this view is supported by the Ks values (supplementary table S3, Supplementary Material online). However, the case of MMDH genes is less certain. In this case, we obtained only a partial CDS assembly of 540 bp (the first 180 codons) for one of the two GgMMDH paralogs of G. gynandra. For this reason, the tree in figure 3F was constructed using partial sequences. In this tree, GgCA MMDH and GgMMDH form a pair and so do MMDHα and ThMMDHα, providing no support for the assumption of being derived from the Th-α WGD. However, the Ks value between ThMMDHα and ThMMDHα is 0.33, so they were likely derived from the Th-α WGD. In comparison, the Ks value between GgC4 MMDH and GgMMDH is only 0.22, substantially smaller than the Gg-α median (0.48). Thus, it is uncertain whether they were derived from the Th-α WGD.

In summary, we propose that all of the eight C4-related genes studied, with the possible exception of MMDH, have retained the Th-α WGD duplicates in G. gynandra and all of them, except for ALAAT, have retained the Th-α WGD duplicates in T. hassleriana.

Upregulation of C4-Related Genes in G. gynandra
The first critical step of the C4 photosynthesis pathway is the conversion of gaseous CO2 to HCO3- in the cytosol of M cells. This reaction is catalyzed by CA (carbonic anhydrase), and HCO3- serves as the substrate for PEPC (phosphoenolpyruvate carboxylase). In G. gynandra, βCA2 and βCA4 are expressed at a much higher level in M cells than in BS cells, suggesting that βCA2 and βCA4 were recruited in C4 photosynthesis (Williams et al. 2016). We predict that AtβCA2, GgβCA2a and ThβCA2a contain a chloroplast transit peptide of 30 amino acids at the first N-terminal by LOCALIZER (Sperschneider et al. 2017), iPSORT (Bannai et al. 2002), and ProtComp 9.0 (a commercial program from Softberry Inc.). This prediction suggests that these proteins are transported to the chloroplast. In contrast, GgC4βCA2a, GgC4βCA2b, and ThβCA2b do not possess the chloroplast transit peptide, suggesting that these βCA2s are expressed in the cytoplasm, and GgC4βCA2a and GgC4βCA2b are much upregulated compared with ThβCA2b (fig. 4A). Note that the C4 form of βCA in C4 Flaveria bidentis was also found to have lost the chloroplast transit peptide and showed an increased expression in the cytosol (Tanz et al. 2009). Therefore, we suggest that GgC4βCA2a and GgC4βCA2b were involved in the evolution of C4 photosynthesis in G. gynandra.

In the case of PEPC paralogs, we found that the upregulated PEPC2, called GgC4 PEPC, in G. gynandra has the C4-type-specific alanine-to-serine change at codon 780 (Christin et al. 2007). In contrast, the other paralogous PEPC is without the alanine-to-serine change and maintains low expression levels in both species (fig. 4C). These data suggest that the upregulated GgC4 PEPC was involved in the C4 evolution in G. gynandra.

Note that GgC4βCA4a, GgC4βCA4b, GgC4NAD-ME, GgC4ALAAT, and GgC4MMDH are also upregulated in G. gynandra (fig. 4).

To test whether the above C4 genes underwent positive selection after duplication, we compared the Ks/Ka ratios between paralogous and orthologous genes (supplementary table S3, Supplementary Material online). Although the Ks/Ka values of the C4 paralogs and orthologs were all smaller than 1, the C4 candidates βCA2, βCA4, PEPC2, NAD-ME2, and ALAAT showed higher nonsynonymous rates than the other paralogous and orthologous genes (supplementary table S3, Supplementary Material online). So, the above C4 enzymes have evolved faster and are expressed dramatically higher in C4 G. gynandra than the corresponding homologous genes in C3 T. hassleriana (fig. 4). Compared with the other Th-α paralogous genes in both species, the C4 enzyme genes have evolved faster than other paralogous non-C4 genes (fig. 3 and supplementary table S3, Supplementary Material online). These observations suggest that the C4 enzyme genes underwent positive selection in C4 G. gynandra.

Unlike the enzymes, the C4 cycle transporters, including TPT, PPT, BASS2, and DIC1, have similar nonsynonymous substitution rates in the two species (supplementary table S3, Supplementary Material online). Importantly, however, one TPT paralog and one PPT paralog are dramatically upregulated in G. gynandra (fig. 4G and H). This observation suggests that these newly evolved transporters are specifically recruited and upregulated for supporting the rapid metabolite shuttles required for the C4 pathway.

Duplicates of Photorespiration Genes in G. gynandra and T. hassleriana
The GLDP catalyzes the decarboxylation of glycolate to release CO2 in mitochondria (Oliver and Raman 1995). A previous study inferred that the GLDP gene duplicates facilitated the establishment of a photorespiratory CO2 pump in C3–C4 intermediate Flaveria species (Schulze et al. 2013). In G. gynandra, there are three GLDP genes: GLDP1, GLDP2a, and GLDP2b (fig. 5A). The Ks value between GLDP2a and GLDP2b is 0.46 (supplementary table S3, Supplementary Material online), so they were apparently derived from the Th-α WGD event. The Ks values between the GLDP1 gene and the GLDP2a and GLDP2b genes are 1.54 and 1.57, respectively, so GLDP1 and the common ancestor of GLDP2a and GLDP2b were apparently derived from the Th-β WGD event. In T. hassleriana, there are only two GLDP genes (GLDP1 and GLDP2) (fig. 5A) and as the Ks value between them is 1.48, they were apparently derived from the Th-β WGD event. Both of these two genes should have been duplicated at the Th-α WGD event, but only one copy of both genes is now found in T.
hassleriana, suggesting loss of a duplicate copy for both genes after the Th-2 WGD. In both G. gynandra and T. hassleriana, the RPKM value of GLDP1 is over 8-fold higher than that of GLDP2 (supplementary fig. S4, Supplementary Material online). This suggests that GLDP1 plays a major role in photorespiration by decarboxylation of glycine in the BS cells of G. gynandra.

ALAAT and GOGAT are involved in critical nitrogen balance for establishing a photorespiratory CO₂ pump at the C₃–C₄ intermediate stage (Mallmann et al. 2014). As
mentioned above, ALAAT has retained the two Th-α duplicates in G. gynandra, though only one copy is retained in T. hasleriana (fig. 3E). There are three GOGAT genes in G. gynandra: GgGOGAT1a, GgGOGAT1b, and GgGOGAT2 (fig. 5B). The Ks value between GgGOGAT1a and GgGOGAT1b is 0.42, so they were likely derived from the Th-α WGD. The Ks values between GgGOGAT2 and GgGOGAT1a (GgGOGAT1b) is 1.30 (1.37) (supplementary table S3, Supplementary Material online), so GgGOGAT2 and the common ancestor of GgGOGAT1a and GgGOGAT1b were apparently derived from the Th-β WGD. In T. hasleriana, only two GOGAT genes are found (fig. 5B) and the Ks value between them is 1.41 (supplementary table S3, Supplementary Material online), so they were likely derived from the Th-β WGD. Thus, T. hasleriana has apparently lost one of two Th-α duplicates.

Our RNA in situ hybridization experiments showed that the GLDP and GOGAT in T. hasleriana, called ThGLDP2 (fig. 6C) and ThGOGAT (fig. 6I), are expressed in BS and M cells, respectively. ThALAAT in T. hasleriana is specifically expressed in M cells (fig. 6O). In G. gynandra, one GLDP paralog, GgGLDP1 (fig. 6A), is mainly expressed in BS cells whereas the other GLDP paralogs, called GgGLDP2a (fig. 6B) and GgGLDP2b (data not shown), are expressed at very low levels in mature leaves. Both GOGAT paralogs in G. gynandra, named GgGOGATa (fig. 6G) and GgGOGATb (fig. 6H), are restricted to express in BS cells. Different from GgGLDP1 and the two GgGLDP paralogs, the C-α-type ALAAT in G. gynandra, called GgC-ALAAT, is expressed in both BS and M cells (fig. 5M). This is not surprising because GgC-ALAAT catalyzes the reversible transfer of an amino group from glutamate to pyruvate, forming alanine and 2-oxoglutarate (2OG) in C-α cycle, to shuttle alanine from BS to M cells in the NAD-ME subtype C-α. G. gynandra (fig. 1). Consistent with this specific C-α mechanism, the other ALAAT paralog, called GgALAATb (fig. 6N), is mainly expressed in the BS cells in G. gynandra for conversion of pyruvate, a product of MAD-ME, to alanine. Alanine shuttled back to M cells is, in turn, converted back to pyruvate in the cytosol and then to PEP by PEPDK in the chloroplast.

From these observations, we suggest that GgGLDP1, GgGOGATa, GgGOGATb, and GgC-ALAAT were recruited to establish a photorespiratory CO2 pump in the BS cells at the C1–C4 intermediate stage during the C4 photosynthesis evolution of G. gynandra. Later, the upregulated C4 CA, PEPC, NAD-MDH, and NAD-ME replaced the photorespiratory CO2 pump mechanism to concentrate CO2 in the BS cells, leading to downregulation of the photorespiration genes in G. gynandra after it evolved C4 photosynthesis. Notably, because GgC-ALAAT is also involved in pyruvate-alanine shuttling in the NAD-ME subtype C4 photosynthesis, its expression level, compared with ThALAAT, is dramatically upregulated not only in BS but also in M cells. We suggest that the GgALAAT duplication played a key role for G. gynandra to enter the C3–C4 intermediate stage during the C4 photosynthesis evolution from C3 photosynthesis.

Not only GLDP and GOGAT but also several other photorespiration genes underwent duplication and retained paralogs in both species. Because the elevated CO2 concentration in BS cells of C4 plants minimizes photorespiration (Sage 2001), those photorespiration genes are repressed in C4 G. gynandra (fig. 7A).

Duplicates of Other Photosynthesis Genes

In plant photosynthesis, the CEF in PSI can create a proton gradient by the cytochrome b6f (Cyt b6f) complex to produce ATPs without production of NADPH (Munekage et al. 2004). Thus, CEF probably contributes to the additional ATPs required for the CCM in C4 photosynthesis. The two CEF pathways in C3 plants are the NDH- and FQR-dependent flows, and the NDH-dependent flow has been suggested to play a central role in C4 photosynthesis (Kubicki et al. 1996; Takabayashi et al. 2005; Ishikawa et al. 2016). Consistent with this notion, our transcriptome analysis revealed that the genes encoding the subunits of the NDH complex are also dramatically upregulated in G. gynandra (fig. 8G). Interestingly, these genes have retained only one copy in G. gynandra, T. hasleriana, and A. thaliana. The reason is unclear, but it could be to maintain a balanced gene dosage. In addition, the expression levels of genes participating in the Cyt b6f complex and FQR-dependent CEF flow are also upregulated in G. gynandra (fig. 8D and H), presumably to boost ATP production. Especially, PetM in the Cyt b6f complex and
**Fig. 6.** In situ hybridization of GLDP, GOGAT, and ALAAT in *Gynandropsis gynandra* and *Tarenaya hassleriana* mature leaves. (A) GgGLDP1 antisense, (B) GgGLDP2a antisense, (C) ThGLDP2 antisense, (D) GgGLDP1 sense, (E) GgGLDP2a sense, (F) ThGLDP2 sense, (G) GgGOGATa antisense, (H) GgGOGATb antisense, (I) ThGOGAT antisense, (J) GgGOGATa sense, (K) GgGOGATb sense, (L) ThGOGAT sense, (M) GgC4ALAAT antisense, (N) GgALAATb antisense, (O) ThALAAT antisense, (P) GgC4ALAAT sense, (Q) GgALAATb sense, and (R) ThALAAT sense. mRNA expression indicated by dark intensity. Bars: 50 μm.
PGR5 in the FQR-dependent CEF flow exhibit upregulation of the Th-α paralogs in G. gynandra.

The NDH complex is associated with PSI by two linker proteins, Lhca5 and Lhca6, to form the NDH–PSI supercomplex for stabilizing the NDH complex (Peng et al. 2009; Peng and Shikanai 2011). Our transcriptome data showed that the PSI and PSI light-harvesting complex (LHCI) genes in G. gynandra display 2 and 1.5 times higher expression levels than in T. hassleriana at the mature leaf stage (S5) (fig. 8E and F). On the other hand, most of the genes encoding photosystem II (PSII), PSII light-harvesting complex (LHCII), and Calvin cycle are expressed at the same or only slightly higher levels in G. gynandra than in T. hassleriana (fig. 8A–C). A previous study on the evolution of photosynthesis genes in Glycine max, Medicago truncatula, and A. thaliana found that PSI, PSII, and LHC genes retain more duplicates derived from WGD but not from single gene duplication because single gene duplication may cause dosage imbalance (Coate et al. 2011). Gynandropsis gynandra and T. hassleriana have retained several Th-α paralogs in PSI, PSII, Cyt b6f, LHCI, LHCII, and Calvin cycle, but these genes show different expression patterns (fig. 8). Several genes in Cyt b6f, PSI, and LHCI derived

### Table A: Photosynthetic genes

| Gene | Gene no. | Gg | Th | S0 | S1 | S2 | S3 | S4 | S5 |
|------|----------|----|----|----|----|----|----|----|----|
| GLDP2 | 2 1 | 0.50| -0.75| -1.03| -1.00| -0.72| -0.94|
| GOGAT | 2 1 | -0.04| -1.07| -1.40| -1.31| -0.78| -1.36|
| GS | 2 2 | -1.28| -1.88| -1.91| -1.45| -1.22| -1.26|
| GOX1 | 2 2 | 0.31| -0.80| -1.11| -0.97| -1.08| -0.93|
| SGAT | 2 1 | -0.82| -2.43| -2.17| -1.86| -1.57| -1.47|
| CATALASE 2 | 3 3 | -0.46| -1.37| -1.62| -1.29| -0.87| -0.80|
| HPR | 1 2 | -1.85| -1.91| -1.45| -0.67| -0.47| -0.37|
| GLYK | 1 2 | -2.42| -3.18| -2.83| -2.35| -2.38| -2.60|
| GGA7 | 1 1 | -0.37| -1.31| -1.30| -0.68| -0.54|
| SMART | 1 1 | -1.59| -1.94| -1.79| -1.32| -1.26| -1.22|
| DIT1 | 2 2 | -0.11| -0.85| -0.90| -0.73| -0.60| -0.54|
| DIT2 | 1 2 | -0.50| -0.72| -0.75| -0.62| -0.61| -0.69|
| BOU | 2 2 | -0.37| -0.98| -1.13| -0.95| -0.66| -0.47|

### Table B: Vein-development related genes

| Gene | Gene no. | Gg | Th | S0 | S1 | S2 | S3 | S4 | S5 |
|------|----------|----|----|----|----|----|----|----|----|
| PIN1 | 2 1 | 1.24| 1.26| 1.24| 0.98| 0.24| -0.59|
| HBI8 | 2 1 | 0.47| 0.30| 0.30| 0.27| 0.15| -0.32|
| MP | 2 2 | -0.39| -0.51| -0.35| -0.70| -0.89| -1.90|
| SCR | 2 2 | -0.50| 0.01| 0.10| -0.06| -0.48| -0.82|
| SHR | 2 2 | -0.53| -0.19| -0.15| 0.11| 0.13| 0.10|
| KANADI 1 | 2 2 | -0.17| 0.08| 0.13| -0.08| 0.18| -0.08|
| KANADI 2 | 2 2 | -1.14| -0.84| -0.87| -0.96| -0.68| -0.67|
| KANADI 4 | 2 2 | 0.77| -0.09| -0.43| -2.19| -4.05| -4.07|
| Dof | 2 2 | -0.09| -0.08| 0.24| 0.16| -0.01| -0.27|
| HBI4 | 1 2 | -0.28| 0.04| -0.08| -0.54| -1.33| -2.00|
| HBI5 | 2 2 | -0.69| -0.79| -0.93| -1.26| -1.60| -2.03|
| XYP1 | 3 3 | -0.14| 0.03| -0.27| -0.99| -2.23| -2.93|
| REVOLUTA | 2 2 | -0.40| -0.34| -0.49| -0.68| -1.02| -1.42|
| APL | 2 2 | 0.19| -0.85| -0.95| -0.43| -0.60| -0.47|

### Table C: Auxin biosynthesis genes

| Gene | Gene no. | Gg | Th | S0 | S1 | S2 | S3 | S4 | S5 |
|------|----------|----|----|----|----|----|----|----|----|
| MYC2 | 2 1 | -1.04| -1.44| -1.66| -1.85| -2.27| -2.41|
| ASBI | 2 1 | 0.59| 0.36| 0.37| 0.26| -0.19| -0.99|
| ASA1 | 1 1 | 1.68| 0.77| 0.33| -0.12| 0.03| -1.17|
| ASA2 | 1 1 | -0.24| 0.24| 0.40| 0.44| 0.51| 0.84|
| PAT1 | 2 1 | 1.44| 1.36| 1.34| 1.24| 0.88| 0.35|
| PAI1 | 1 1 | -0.44| 0.10| 0.11| -0.19| -0.78| -1.12|
| GIPS1 | 2 1 | 0.43| 1.25| 1.24| 1.03| 0.69| 0.07|
| GIPS2 | 1 1 | 1.22| 0.88| 0.68| 0.96| 1.25| 1.52|
| TSA2 | 1 1 | -0.24| 0.24| 0.60| 0.51| 0.38| 0.43|
| TSA1 | 1 1 | -0.33| 0.01| 0.14| 0.52| 0.68| 0.51|
| TSB2 | 1 1 | 1.15| 1.25| 1.30| 1.57| 1.74| 1.52|
| NIT4 | 1 1 | -0.28| -0.20| -0.12| -0.05| 0.21| 0.49|
| TAA | 2 1 | -1.22| -0.69| -1.00| -0.82| -1.47| -2.08|
| TAR | 1 1 | 0.66| 1.25| 0.74| 0.32| 0.29| 0.36|
| YUC2 | 1 1 | -0.32| -0.19| 0.14| 0.43| -0.88| 1.82|
| YUC4 | 1 1 | -0.38| -1.65| -1.38| -1.35| -1.38| -1.11|
| YUC6 | 1 1 | -0.32| -0.19| 0.14| 0.43| -0.88| 1.82|
| AMII | 1 1 | -0.62| -0.75| -0.56| -0.39| -0.22| 0.14|

Fig. 7. Expression ratios of genes and numbers of paralogous genes involved in (A) photorespiration, (B) vein development, and (C) auxin biosynthesis. The first column shows gene names. The second and third columns show paralogous gene numbers in Gynandropsis gynandra and Tarenaya hassleriana. The fourth to ninth columns show gene expression ratios between G. gynandra and T. hassleriana. S0–S5 denote the six leaf developmental stages. The color bar indicates the fold differences (log2 ratios) in gene expression between G. gynandra and T. hassleriana. Gene no.: gene number.
from the Th-x WGD show increased expression in *G. gynandra* compared with *T. hassleriana*. Although several Th-x paralogs in PSI, LHCII, and Calvin cycle are retained, most paralogs still maintain similar expression levels in both of these C₃ and C₄ species. Therefore, the observation of Th-x paralogs showing increased expression levels in PSI is

---

**Figure 8.** Expression ratios of genes and paralogous gene numbers involved in photosynthesis. (A) Light-harvesting complex II. (B) PSII. (C) Calvin cycle. (D) Cytochrome b6f complex. (E) Light-harvesting complex I. (F) PSI. (G) NDH-dependent CEF flow. (H) FQR-dependent CEF flow. In each panel, the first column shows the gene names. The second to fourth columns show paralogous gene numbers in *Gynandropsis gynandra*, *Tarenaya hassleriana*, and *Arabidopsis thaliana*. The fifth to tenth columns show the expression ratios of genes between *G. gynandra* and *T. hassleriana*. The color bar indicates the fold differences (log2 ratios) in gene expression between *G. gynandra* and *T. hassleriana* at the same stage. Gene no.: gene number.
consistent with the view that upregulated subunits of the NDH complex is to stabilize the NDH complex. Additionally, LHCI subunits play a crucial role for light harvesting in PSI–LHCI supercomplex (Bressan et al. 2016). We suggest that the Th-α paralogs with upregulation in LHCI contribute to increased absorption of photons by the PSI–LHCI supercomplex to transfer more excitation energy to CEF. Then, the upregulated Cyt b6f complex could transfer more electrons to increase proton pumping by enhanced CEF due to upregulated NDH- and FQR-dependent flows, which produce extra ATPs for CCM in the C4 G. gynandra.

Duplicates of Leaf Vein Development-Related Genes

Examining the Th-α duplicate gene pairs in auxin biosynthesis pathways, we found that IGPS and TSA1 have retained duplicates in T. hassleriana, whereas ASB1 and PAT1 have retained duplicates in G. gynandra (fig. 7C). In G. gynandra, both ASBi1and PAT1 are upregulated over 1.5 and 2.5 times during early leaf development (S0), compared with T. hassleriana (fig. 7C). We also found that MYC2, a negative regulator of auxin biosynthesis (Dombrecht et al. 2007), has retained two duplicated genes in T. hassleriana but only one copy in G. gynandra, leading to a higher expression of MYC2 in T. hassleriana than G. gynandra. In our previous study, a lower MYC2 expression in G. gynandra resulted in higher auxin biosynthesis than in T. hassleriana (Huang et al. 2017), supporting our hypothesis that increased auxin level and transport is required for developing high vein density, an important feature of Kranz leaf anatomy in C4 plants.

We also investigated genes involved in vein development (Huang et al. 2017) in the two species. In both species, almost all known vein-development-related genes have retained their duplicates, including MONOPTEROS, Homeobox gene 15 (HB15), HB14, HB8, SHR, SCR, PIN1, REVOLUTA, XYP1, APL, Dof-type zinc finger (AT2G28510), KANADI1 (KAN1), KAN2, and KAN4 (fig. 7B). Most of these genes are expressed at similar levels in the two species during early leaf development (S0–S2, fig. 7B). The prevailing retention of the vein-development-related duplicated genes probably has contributed to vein complexity in the two species, that is, septenary order venation in G. gynandra and senary order venation in T. hassleriana whereas only quinary order venation in A. thaliana (Huang et al. 2017). Additionally, the early vein-development-related genes, HB8 and PIN1, have retained duplicated copies and are expressed at higher levels in G. gynandra (fig. 7B). Thus, it might be the additional copies and the increased expression level of PIN1 and HB8, combined with low expression of MYC2, that have led to the higher vein density in G. gynandra than in T. hassleriana.

Discussion

Gene duplication, either whole genome or single gene duplication, is considered a precondition for C4 evolution (Sage 2004). Several C4 enzyme genes, including PEPC, PPDK, NADP-ME, NADP-MDH, and CA have been duplicated and then became involved in the evolution of C4 photosynthesis in the genus Flavera (Monson 2003). Although gene duplication has long been thought to be important in C4 photosynthesis evolution, the focus so far has been on the modification of C4 enzyme genes. Our study suggests that the most recent WGD, Th-α, in the Cleomaceae played an important role not only in C4 cycle formation but also in vein patterning and the establishment of a photorespiratory CO2 pump at the C3–C4 intermediate stage.

During C4 evolution at the C3–C4 intermediate stage, restriction of the GLDP expression to BS cells was an important step to establish a photorespiratory CO2 pump after GLDP duplication in Flaveria (Hylton et al. 1988; Schulze et al. 2016). Our study showed that not only GLDP but also GOGAT and ALAAT have retained Th-α WGD paralogs in G. gynandra, a C4 plant with NAD-ME subtype pathway. Importantly, RNA in situ hybridization experiments showed that GgGLDP1, GgGOGATa, and GgGOGATb are restricted to express in BS cells. Moreover, GgC4ALAAT is also restricted to express in BS cells. Therefore, we suggest that the Th-α WGD facilitated the establishment of a photorespiratory CO2 pump at the C3–C4 intermediate stage and the maintenance of nitrogen balance during C4 photosynthesis formation in G. gynandra.

Tarenaya hassleriana has retained ThGLDP paralogs but lost ThGOGAT and ThALAAT duplicates. RNA in situ hybridization experiments showed that ThGLDP1 and ThGOGAT are expressed in both BS and M cells, and ThALAAT is expressed only in M cells. Thus, T. hassleriana failed to establish a photorespiratory CO2 pump. We suggest that losses of ThALAAT and ThGOGAT paralogs were the reason why T. hassleriana failed to evolve into the C3–C4 intermediate stage.

A previous analysis of the sorghum (the NADP-ME subtype C4 plant) genome concluded that the WGD duplicated copies of PEPC and NADP-ME have been preserved, whereas other C4 enzyme gene paralogs were probably lost (Wang et al. 2009). In this study, we found six or five genes in G. gynandra and five genes in T. hassleriana encoding C4 cycle enzymes have retained duplicated copies after the Th-α WGD (fig. 3). Although the C4 cycle genes that were duplicated, including βCA2, βCA4, PEPC2, NAD-ME2, and ALAAT, showed no case of Kα/Kβ>1, they showed higher nonsynonymous rates than their paralogous and orthologous genes. Thus, it is possible that these C4 cycle enzymes only need to change some important amino acids, such as the specific alanine-to-serine transition in C4 PEPC (Bailey and Elkan 1994), to alter their catalytic properties for functioning in the C4 cycle after gene duplication. However, the duplicates of the four C4 cycle transport genes, BASS2, DIC1, PTP, and PPT, have similar nonsynonymous substitution rates between their homologs (supplementary table S3, Supplementary Material online), probably because the transporters only function in moving C4 cycle intermediates but not in enzyme reaction. Importantly, the predicted C4 cycle genes are dramatically upregulated in G. gynandra (fig. 4). Thus, it is possible that these C4 cycle genes underwent a short period of positive selection and increased their expression levels for recruitment into the C4 cycle after the Th-α WGD event.

CEF has been suggested to generate additional ATPs for the C4 CCM (Munekage et al. 2004). Two distinct CEF pathways, NDH- and FQR-dependent flows, have been identified.
in C₃ plants. Both pathways transfer electrons to the Cyt b₆f complex, creating a greater proton gradient across the thylakoid membrane of chloroplasts, which is then used to drive ATP synthesis. The NDH-dependent flow was shown to play a role in C₄ photosynthesis (Takabayashi et al. 2005; Ishikawa et al. 2016), and in Flaveria both pathways showed higher activities in C₄ species than in C₃ species (Nakamura et al. 2013). Our study also finds that both pathways, especially the NDH-dependent flow, are upregulated in C₄ G. gynandra (fig. 8G and H). In addition, the genes encoding Cyt b₆f complex subunits are upregulated in G. gynandra (fig. 8D), which may allow the Cyt b₆f complex to accept more electrons to create additional ATPs for the C₄ cycle. The association of the NDH complex with PSI through Lhca5 and Lhca6 to form the NDH–PSI supercomplex (Peng et al. 2009; Peng and Shikanai 2011) may enhance the CEF function. We also found that the genes encoding PSI and LHCl are upregulated in G. gynandra (fig. 8E and F), which may enhance CEF in G. gynandra.

Compared with the expression levels of photosynthesis genes in both species, the genes encoding the Cyt b₆f complex, LHCl, PSI, and CEF proteins are upregulated in G. gynandra (fig. 8D–H), whereas the genes encoding PSII, LHCl, and Calvin cycle proteins are not. Therefore, the CEF-associated complex might play an important role in generating extra ATPs in C₄ G. gynandra.

The chloroplast NDH complex is encoded by 11 subunit genes in the plastid genome and 20 subunit genes in the nuclear genome (Shikanai 2016). Interestingly, there are no Th-α paralogs for any of the 20 nuclear-encoded subunits in both species and A. thaliana, suggesting that there was no advantage for retaining the duplicates for these genes. In contrast, several genes participating in photosynthesis have retained their Th-α duplicates. Moreover, those paralogous genes encoding the Cyt b₆f complex subunits, LHCl, PSI, and FQR-dependent flow exhibit higher expression levels in G. gynandra than in T. hassleriana (fig. 8D–F and H). It seems that the Th-α WGD facilitated the upregulation of CEF to capture more protons and produce additional ATPs for the C₄ photosynthesis CCM.

Finally, we found that almost all known vein-development-related genes have retained their duplicates after the Th-α WGD (fig. 7B). We suggest that the vein-development-related gene duplicates have contributed to more complex leaf venation architecture and probably differentiation of M-related gene duplicates have contributed to more complex ATPs in C₄ complex might play an important role in generating extra Calvin cycle proteins are not. Therefore, the CEF-associated complex, creating a greater proton gradient across the thylakoid membrane of chloroplasts, which is then used to drive ATP synthesis. The NDH-dependent flow was shown to play a role in C₄ photosynthesis (Takabayashi et al. 2005; Ishikawa et al. 2016), and in Flaveria both pathways showed higher activities in C₄ species than in C₃ species (Nakamura et al. 2013). Our study also finds that both pathways, especially the NDH-dependent flow, are upregulated in C₄ G. gynandra (fig. 8G and H). In addition, the genes encoding Cyt b₆f complex subunits are upregulated in G. gynandra (fig. 8D), which may allow the Cyt b₆f complex to accept more electrons to create additional ATPs for the C₄ cycle. The association of the NDH complex with PSI through Lhca5 and Lhca6 to form the NDH–PSI supercomplex (Peng et al. 2009; Peng and Shikanai 2011) may enhance the CEF function. We also found that the genes encoding PSI and LHCl are upregulated in G. gynandra (fig. 8E and F), which may enhance CEF in G. gynandra.

Compared with the expression levels of photosynthesis genes in both species, the genes encoding the Cyt b₆f complex, LHCl, PSI, and CEF proteins are upregulated in G. gynandra (fig. 8D–H), whereas the genes encoding PSII, LHCl, and Calvin cycle proteins are not. Therefore, the CEF-associated complex might play an important role in generating extra ATPs in C₄ G. gynandra.

The chloroplast NDH complex is encoded by 11 subunit genes in the plastid genome and 20 subunit genes in the nuclear genome (Shikanai 2016). Interestingly, there are no Th-α paralogs for any of the 20 nuclear-encoded subunits in both species and A. thaliana, suggesting that there was no advantage for retaining the duplicates for these genes. In contrast, several genes participating in photosynthesis have retained their Th-α duplicates. Moreover, those paralogous genes encoding the Cyt b₆f complex subunits, LHCl, PSI, and FQR-dependent flow exhibit higher expression levels in G. gynandra than in T. hassleriana (fig. 8D–F and H). It seems that the Th-α WGD facilitated the upregulation of CEF to capture more protons and produce additional ATPs for the C₄ photosynthesis CCM.

Finally, we found that almost all known vein-development-related genes have retained their duplicates after the Th-α WGD (fig. 7B). We suggest that the vein-development-related gene duplicates have contributed to more complex leaf venation architecture and probably differentiation of M and B5 cells in G. gynandra (septenary orders) and T. hassleriana (senary orders) than in A. thaliana (quinary orders) (Huang et al. 2017). Unlike the vein-development-related genes, only a few genes in auxin biosynthesis pathways have retained duplicates, but most of these genes are upregulated in G. gynandra (fig. 7C). It might be that only some steps in the auxin biosynthesis pathway may be rate limiting and require higher dosages of the biosynthetic enzymes to boost the formation of high vein density in C₄ leaves. Interestingly, MYC2, a negative regulator of tryptophan biosynthesis, has lost its duplicate in G. gynandra, leading to a low expression level in G. gynandra and thus a higher level of auxin biosynthesis for vein development. In addition, an important auxin transport gene, PIN1, and an early vein-development related gene, HB8, have retained duplicates in G. gynandra but not in T. hassleriana, resulting in higher PIN1 and HB8 expression levels in G. gynandra. Thus, the low expression of the negative regulator MYC2 for auxin biosynthesis coupled with the additional copies and the increased expression levels of PIN1 and HB8 in G. gynandra together have likely contributed to the high vein density in the leaves of G. gynandra (Huang et al. 2017).

In summary, our study has provided evidence that the Th-α WGD facilitated the C₄ photosynthesis evolution in G. gynandra, including the early step of increased vein density, the establishment of a photosynthetic CO2 pump at the C₃–C₄ intermediate stage, the recruitment of C₄ cycle genes in C₄ cycle and upregulated CEF-associated complex for production of extra energy for the C₄ cycle. Although C₃ T. hassleriana shared the Th-α WGD with G. gynandra, it has stayed at the anatomical preconditioning stage of increased vein density (senary order venation) compared with quinary order venation in A. thaliana, likely because it did not undergo the C₅–C₄ intermediate stage for establishing the photosynthetic CO2 pump.

Materials and Methods

Plant Material and RNA Isolation

Gynandropsis gynandra was grown in growth chambers under the light-dark cycle: 12 h light (200–250 μmol m⁻² s⁻¹) at 27 °C and 12 h darkness at 25 °C. For isolation of total RNA, fifth leaves (~0.5–10 mm long) of 20 plants (12–14 days old) were harvested at midday and immediately frozen in liquid nitrogen. Total RNA was isolated by TRIzol reagent (Invitrogen, Carlsbad, CA, USA) according to the manufacturer’s instructions, using a 1:2 ratio of sample:TRIzol reagent. RNA samples were treated with DNase I at 37 °C for 30 min to eliminate contaminating genomic DNA.

Iso-Seq cDNA Library Preparation and SMRT Sequencing

The total RNA sample of G. gynandra leaves was used to construct a cDNA library by the High Throughput Genomics Core, Academia Sinica, Taiwan. Full-length cDNAs were synthesized from HQ total RNA with oligo-dT priming, and amplified using the Clontech SMARTer PCR cDNA Synthesis Kit with PrimeSTAR GXL DNA Polymerase (Takara Bio, USA). The PCR products were purified with AMPure PB magnetic beads (Pacific Biosciences, USA). The library profiles showed distribution from 300 bp to 6 kb, with major peaks at around 1.8 kb. To minimize loading bias, two size-fractionated libraries (≤2 kb and >2 kb) were isolated on 0.75% gel cassette using the Qubit dsDNA HS Assay Kit with the Qubit 3.0 Fluorometer (Thermo Fisher Scientific, USA) and Agilent BioAnalyzer 2100 High Sensitivity DNA Kit (Agilent
Technologies, USA), respectively. Finally, the libraries were pooled and loaded into the SMRTcell and sequenced on the PacBio Sequel platform by the High Throughput Genomics Core, Academia Sinica, Taiwan.

**PacBio Long-Read Processing**

The PacBio raw reads were classified into CCS and non-CCS subreads by running the IsoSeq3 module in PacBio SMRT Analysis v6.0. IsoSeq v3 determines a CCS or subread sequence to be full length if the 5\(^\prime\)- and 3\(^\prime\)-primers and poly A tail signal were all present in the correct order. After clustering, the full-length transcripts were polished with predicted consensus accuracy \(>0.99\), which were considered polished HQ transcripts. To correct the potential indel errors, we polished the full-length transcripts using the trimmed Illumina RNA-seq reads deposited at NCBI (Huang et al. 2017) as input to the Pilon software (Walker et al. 2014).

**Assembly of RNA-Seq Reads and Construction of ORF Databases**

The Illumina RNA-seq reads from *G. gynandra* and *T. hassleriana* deposited at NCBI (Huang et al. 2017) were trimmed using the Trimmomatic tool (Bolger et al. 2014). After the quality trimming at Q30, the trimmed reads were end merged to generate longer reads by FLASH (Magoc and Salzberg 2011). The merged and unmerged paired-end reads in each species were assembled de novo, using the CLC Genomics Workbench (QIAGEN, Germany) with default options.

To construct the ORF database of *G. gynandra*, we used the following procedure (supplementary fig. S1, Supplementary Material online): First, the PacBio polished HQ isoform transcripts were combined with the CLC assembled contigs for predicting CDSs by the orf-finder-py tool (Stewart et al. 2017). Redundant CDSs were removed by CD-HIT (Li and Godzik 1999; Fu et al. 2012) and the longest CDSs in the combined database were retained as the representative transcripts. To annotate the CDSs, we used BLASTp (Camacho et al. 2009) against the Araport11 gene database (Camacho et al. 2009). CDSs in the representative transcripts covering over 90% of the length of their target *A. thaliana* genes (Araport11) (Cheng et al. 2017) were collected to form a full-length CDS data set. Second, we mapped the Illumina RNA-seq reads against the full-length CDSs using Bowtie 2 (Langmead and Salzberg 2012). Then, we collected the unmapped reads for de novo assembly using CLC. The new full-length CDSs from the second assembly were added to the full-length CDS data set. Third, to obtain more full-length CDSs, the remaining transcripts (<90% coverage against the target *A. thaliana* genes) from PacBio iso-seq and CLC assembled contigs were used with the assembled CDSs using CAP3 (Huang and Madan 1999). Finally, we added the CAP3 assembly to the full-length CDS data set and removed redundant CDSs. The final *G. gynandra* CDS data set contained 21,535 ORFs in which 16,535 ORFs were full-length CDSs.

The procedure for constructing the *T. hassleriana* CDS database was similar to the above. The previously constructed *T. hassleriana* CDSs (Cheng et al. 2013), which were inferred from the annotated genes of *T. hassleriana*, were combined with the CDSs from the CLC-assembled contigs. The remaining steps were the same as the procedure of the construction of the *G. gynandra* CDS database (supplementary fig. S1, Supplementary Material online). The final *T. hassleriana* CDS database contained 27,617 CDSs with 22,511 full-length CDSs.

**Estimating Gene Expression Levels**

To quantify the expression levels of the assembled CDSs in a species (*G. gynandra* or *T. hassleriana*), the Illumina reads of the six developmental stages (S0–S5) (Külahoğlu et al. 2014) deposited at NCBI were subjected to quality trimming at Q30, and were mapped to the corresponding ORF database for that species. The single-end read data were then mapped to the ORFs using Bowtie 2 (Langmead and Salzberg 2012) with default settings. Finally, the eXpress software (Roberts and Pachter 2013) was used to resolve the multihit reads and calculate the relative measurements of RPKMs as the expression levels of the CDSs.

To have meaningful comparisons of gene expression levels for the six developmental stages between two species, the RPKM values were normalized using the upper quartile normalization procedure (Bullard et al. 2010), using the S0 stage in *G. gynandra* as the reference.

**In Situ Hybridization**

In situ hybridization experiments were carried out as described by Jackson (Jackson 1991). Plant material was fixed in 4% paraformaldehyde (Sigma) in 0.1 M phosphate buffer (pH 7.2) for 16 h at 4 °C and embedded in Paraplast Plus (Sigma-Aldrich). Sections (12 μm) were cut using a microtome (RM 2135; Leica), and collected in xylene-coated slides. Slides were deparaffinized and treated with 20 mg/ml proteinase K. In vitro transcription of the digoxigenin UTP-labeled (Roche) RNA sense and antisense probes were obtained using T7 and Sp6 polymerases. Primers used to generate the probe clones are listed in supplementary table S4, Supplementary Material online. Hybridization was performed in hybridization solution at 50–55 °C overnight. Digoxigenin detection and signal visualization were carried out using nitroblue tetrazolium and 5-bromo-4-chloro-3-indolyl phosphate (Roche), following the manufacturer’s instructions. Slides were air dried and mounted with Kaiser’s glycerol gelatine mounting medium (Sigma-Aldrich).

**Supplementary Material**

Supplementary data are available at Molecular Biology and Evolution online.

**Acknowledgments**

This study was supported by Academia Sinica (AS-TP-109-L10). We thank the High-Throughput Sequencing Core, Biodiversity Research Center, Academia Sinica for high-throughput sequencing and support.
Author Contributions
C.-F.H. and W.-H.L. designed the study. W.-H.L. supervised the study. C.-F.H. conducted bioinformatics analyses. C.-F.H. and W.-Y.L. performed in situ hybridization experiments. M.-Y.J.L., and Y.-H.C. performed PacBio sequencing. C.-F.H. and Y.-H.C. captured microscopic images. C.-F.H. and W.-H.L. wrote the first draft. C.-F.H., W.-Y.L., M.-Y.J.L., Y.-H.C., M.S.B.K., and W.-H.L. wrote the article.

Data Availability
The sequence data have been deposited in the Sequence Read Archive, www.ncbi.nlm.nih.gov/sra (SRR13973106, SRR13973107, and SRR13973108) and the Transcriptome Shotgun Assemblies, www.ncbi.nlm.nih.gov/Traces/wgs/?view=TSA (GJBA00000000 and GFM02000000).

References
Bailey TL, Elkan C. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol. 2:28–36.
Bannai H, Tamada Y, Maruyama O, Nakai K, Miyano S. 2002. Extensive analyses of the Cleome transcriptome elucidate the history of genome duplications in Arabidopsis and other brassicas. Genome Biol Evol. 1:391–399.
Bauwe H. 2011. Photorespiration: the bridge to C4 photosynthesis. In: Barker MS, Vogel H, Schranz ME. 2009. Paleopolyploidy in the Brassicales: analyses of the Cleome transcriptome elucidate the history of genome duplications in Arabidopsis and other brassicas. Genome Biol Evol. 1:391–399.
Bauwe H. 2011. Photorespiration: the bridge to C4 photosynthesis. In: Raghavendra AS, Sage RF, editors. C4 photosynthesis evolution: modular, individually adaptive steps on a Mount Fuji fitness landscape. Cell 153(7):1579–1588.
Bower JE, Chapman BA, Rong J, Paterson AH. 2003. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature 422(6930):433–438.
Bressan M, Dall’Osto L, Bargigia I, Alcocer MJP, Viola D, Cerullo G, D’Andrea C, Bassi R, Ballottari M. 2016. LHCII can substitute for LHCI as an antenna for photosystem I but with reduced light-harvesting capacity. Nat Plants 2:16131.
Bullard JH, Purdom E, Hansen KD, Dudoit S. 2010. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics. 11:94–94.
Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. BMC Bioinformatics. 10:421.
Cheng CY, Krishnakumar V, Chan AP, Thibaud-Nissen F, Schobel S, Town CD. 2017. Araport11: a complete reannotation of the Arabidopsis thaliana reference genome. Plant J. 89(4):789–804.
Cheng S, van den Berg E, Zeng P, Zhong X, Xu J, Liu X, Hofberger J, de Bruijn S, Bhide AS, Kuo SH, Guo Y, et al. 2013. The Tarenaya hassleri genome provides insights into reproductive trait and genome evolution of crucifers. Plant Cell 25(8):2813–2830.
Christin P-A, Osborne CP, Chatellet DS, Columbus JT, Besnard G, Hodkinson TR, Garrison LM, Vorontsova MS, Edwards EJ. 2013. Anatomical enablers and the evolution of C4 photosynthesis in grasses. Proc Natl Acad Sci U S A. 110(4):1381–1386.
Christin PA, Salamin N, Savolainen V, Duvall MR, Besnard G. 2007. C4 Photosynthesis evolved in grasses via parallel adaptive genetic changes. Curr Biol. 17(14):1241–1247.
Coate JE, Schlueter JA, Whaley AM, Doyle JJ. 2011. Comparative evolution of photosynthetic genes in response to polyploid and nonpolyploid duplication. Plant Physiol. 155(4):2081–2095.
Dombret D, Xue GP, Sprague SJ, Kirkegaard JA, Ross JJ, Reid JB, Fitt GP, Sewelam N, Schenk PM, Manners J, et al. 2007. MYC2 differentially modulates diverse jasmonate-dependent functions in Arabidopsis. Plant Cell 19(7):2225–2245.
Emms DM, Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 20(1):238.
Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28(23):3150–3152.
Gowik U, Westhoff P. 2011. The path from C3 to C4 photosynthesis. Plant Physiol. 155(1):56–63.
Hatch MD. 1987. C4 photosynthesis: a unique belond of modified biochemistry, anatomy and ultrastructure. Biochim Biophys Acta. 895(2):81–106.
Heckmann D, Schulze S, Denton A, Gowik U, Westhoff P, Weber AP, Lercher MJ. 2013. Predicting C4 photosynthesis evolution: modular, individually adaptive steps on a Mount Fuji fitness landscape. Cell 153(7):1579–1588.
Huang CF, Yu CP, Wu YH, Lu M, Tu SL, Wu SH, Shiu SH, Ku MSB, Li WH. 2017. Elevated auxin biosynthesis and transport underlie high vein density in C4 leaves. Proc Natl Acad Sci U S A. 114(33):8688–8691.
Huang X, Madan A. 1999. CAP3: a DNA sequence assembly program. Genome Res. 9(9):868–877.
Hytonen AM, Rawsthorne SM, Smith AM, Jones DA, Woolhouse HW. 1988. Glycine decarboxylase is confined to the bundle-sheath cells of leaves of C3-C4 intermediate species. Planta 175(4):452–459.
Ifuku K, Endo T, Shikanai T, Aro EM. 2011. Structure of the plastocyanin NDH dehydrogenase-like complex: nomenclature for nuclear-encoded subunits. Plant Cell Physiol. 52(9):1560–1568.
Ishikawa N, Takabayashi A, Noguchi K, Tazoe Y, Yamamoto H, von Caemmerer S, Sato F, Endo T. 2016. NDH-mediated cyclic electron flow around photosystem I is crucial for C4 photosynthesis. Plant Cell Physiol. 57(10):2020–2028.
Jackson D. 1991. In situ hybridization in plants. Molecular plant pathology: a practical approach. In: Gurr SJ, McPherson MJ, Bowles DJ, editors. London, UK: Oxford University Press. p. 163–174.
Kubicki A, Funk E, Westhoff P, Steinmüller K. 1996. Differential expression of plastome-encoded ndh genes in mesophyll and bundle-sheath chloroplasts of the C4 plant Sorghum bicolor indicates that the complex I-homologous NAD(P)H-plastoquinone oxidoreductase is involved in cyclic electron transport. Planta 199(2):276–281.
Kulahoglu C, Denton AK, Sommer M, Ma6 S, Schliesky S, Wrobel T, Berckmans B, Gongora-Castillo E, Buell CR, Simon R, et al. 2014. Comparative transcriptome atlases reveal altered gene expression modules between two Cleomaceae C3 and C4 plant species. Plant Cell. 26(8):3243–3260.
Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods. 9(4):357–359.
Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22(13):1658–1659.
Monosekage Y, Hashimoto M, Miyake C, Tomizawa K, Endo T, Tasaka M, Shikanai T. 2004. Cyclic electron flow around photosystem I is essential for photosynthesis. Nature 429(6991):579–582.
Monson RK. 1999. The origins of C4 genes and evolutionary pattern in the C4 metabolic phenotype. In: Sage RF, Monson RK, editors. C4 photosynthesis and related processes: chemistry, anatomy and ultrastructure. Biochim Biophys Acta. 895(2):81–106.
Madden TL. 2009. BLAST: architecture and applications. BMC Bioinformatics. 10:421.
