Gene and genome duplications and the origin of C4 photosynthesis: Birth of a trait in the Cleomaceae

Erik van den Bergh a, Canan Külahoglu b, Andrea Bräutigam b, Julian M. Hibberd c, Andreas P.M. Weber b, Xin-Guang Zhu d, M. Eric Schranz a,∗

a Biosystematics, Wageningen University and Research, Drenthendaalweg 1, 6708 PB Wageningen, The Netherlands
b Institute of Plant Biochemistry, Center of Excellence on Plant Sciences(CEPLAS), Heinrich-Heine-University, D-40225 Düsseldorf, Germany
c Department of Plant Sciences, University of Cambridge, Cambridge CB2 3EA, United Kingdom
d Plant Systems Biology Group, Partner Institute of Computational Biology, Chinese Academy of Sciences/Max Planck Society, Shanghai 200031, China

ARTICLE INFO
Article history:
Received 15 May 2014
Received in revised form 19 August 2014
Accepted 23 August 2014

Keywords:
Plant genome evolution
Synteny
Cleomaceae
Brassicaceae
Bioinformatics
Whole genome duplication
Paleopolyploidy
C4 photosynthesis

ABSTRACT

C4 photosynthesis is a trait that has evolved in 66 independent plant lineages and increases the efficiency of carbon fixation. The shift from C3 to C4 photosynthesis requires substantial changes to genes and gene functions effecting phenotypic, physiological and enzymatic changes. We investigate the role of ancient whole genome duplications (WGD) as a source of new genes in the development of this trait and compare expression between paralog copies. We compare Gynandropsis gynandra, the closest relative of Arabidopsis that uses C4 photosynthesis, with its C3 relative Tarenaya hassleriana that underwent a WGD named Th-α. We establish through comparison of paralog synonymous substitution rate that both species share this paleohexaploidy. Homologous clusters of photosynthetic gene families show that gene copy numbers are similar to what would be expected given their duplication history and that no significant difference between the C3 and C4 species exists in terms of gene copy number. This is further confirmed by synteny analysis of T. hassleriana, Arabidopsis thaliana and Aethionema arabicum, where syntenic region copy number ratios lie close to what could be theoretically expected. Expression levels of C4 photosynthesis orthologs show that regulation of transcript abundance in T. hassleriana is much less strictly controlled than in G. gynandra, where orthologs have extremely similar expression patterns in different organs, seedlings and seeds. We conclude that the Th-α and older paleopolyploidy events have had a significant influence on the specific genetic makeup of Cleomaceae versus Brassicaceae. Because the copy number of various essential genes involved in C4 photosynthesis is not significantly influenced by polyploidy combined with the fact that transcript abundance in C. gynandra is more strictly controlled, we also conclude that recruitment of existing genes through regulatory changes is more likely to have played a role in the shift to C4 than the neofunctionalization of duplicated genes.

DATA: The data deposited at NCBI represents raw RNA reads for each data series mentioned: 5 leaf stages, root, stem, stamen, petal, carpel, sepal, 3 seedling stages and 3 seed stage stages of Tarenaya hassleriana and Gynandropsis gynandra. The assembled reads were used for all analyses of this paper where RNA was used. http://www.ncbi.nlm.nih.gov/Traces/sra/?study=SRP036637, http://www.ncbi.nlm.nih.gov/Traces/sra/?study=SRP036837

© 2014 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/3.0/).

1. Introduction

Over sixty lineages of both monocot and eudicot angiosperms have evolved a remarkable solution to maximize photosynthesis efficiency under low CO2 levels, high temperatures and/or drought: C4 photosynthesis [1]. The evolution of this modified photosynthetic pathway represents a wonderful example of convergent evolution. While the changes necessary for the transition from C3 to C4 photosynthesis are numerous, the trait has a wide phylogenetic distribution across angiosperms, with 19 different plant
families across the globe to contain one or multiple members capable of C₄ photosynthesis [2]. Much research on eudicot C₄ has focused on Flaveria species (Asteraceae), which contains not only C₄ species but also a number of C₃/ C₄ intermediates [3]. With the emergence of genomics and the choice of Arabidopsis thaliana as the genomics standard model organism, species in the Cleomaceae, a sister-family to the Brassicaceae (containing Arabidopsis and Brassica crops) have been proposed for genetic studies of C₄ [4,5].

C₄ plants spatially separate the fixation of carbon away from the Rubisco active site by using phosphoenolpyruvate carboxylase, an alternate carboxylase that does not react with oxygen. As a consequence they are more efficient under permissive conditions [6]. The typical C₄ system is characterized by a morphological change: so-called Kranz anatomy [7]. In this anatomy, specialized mesophyll (M) cells surround enlarged bundle sheath (BS) cells, with the leaf veins internal to the BS. Generally, the veination in C₄ leaves is increased [8]. This internal leaf architecture physically partitions the biochemical events of the C₄ pathway into two main phases. In the first phase, dissolved HCO₃⁻ is assimilated into C₄ acids by phosphoenolpyruvate carboxylase (PEPC) in the mesophyll cells. In the second phase, these acids diffuse into the chloroplast loaded bundle sheath (BS) cells, where they are decarboxylated and the released CO₂ is fixed by Rubisco. The increased CO₂ concentration in the BS cells allows carbon fixation by Rubisco to be much more efficient by reducing photorespiration. Two subtypes of the C₄ biochemical pathway are defined, based on the most active C₄ acid decarboxylase that liberates CO₂ from C₄ acids in the bundle sheath: NADP-malic enzyme (NADP-ME), NAD-malic enzyme (NAD-ME); a facultative addition of phosphoenolpyruvate carboxykinase (PEPCK) activity can be present in either subtype [9]. The subtypes are used as a classification scheme for C₄.

The process of carboxylation and decarboxylation costs more energy than the simpler C₃ form of photosynthesis, but it diminishes photorespiration. In conditions of low atmospheric CO₂ pressure, photorespiration causes a major loss in photosynthetic output and the elaborate concentrating mechanisms of C₄ photosynthesis circumvent this [10].

All genes important for the C₄ pathway are expressed at relatively low levels in C₃ leaves [11]. The mechanism for recruitment of these genes into the C₄ pathway remains to be elucidated. For some ancestral C₃ genes changes in cis-regulatory elements, while in others changes in trans generate M and BS cell specificity [12–14], indicating variation in the mechanisms underlying gene recruitment into the C₄ pathway. It has been proposed that gene duplication and subsequent neo-functionalization of one gene copy has facilitated the alterations in gene expression that underlie the evolution of C₄ photosynthesis [15,16]. Gene duplication is proposed to be a (pre)condition for the evolution of C₄ because it allows the organism to maintain the original gene while a duplicate version can acquire beneficial changes. This can lead to significant changes in metabolism without the deleterious effect of modifications to essential genes. A recent study that compared convergent evolution of photosynthetic pathways with parallel evolution concluded that duplications are not essential for the development of C₄ biochemistry, but rather changes in expression and localization of specific genes [11,17]. However, this study highlighted just the number of C₄ genes and did not take into account the age and mechanism of gene duplications.

The modifications necessary for the anatomical changes from C₃ to C₄ photosynthesis are not well established. Recent work has shown that the SCARECROW (SCR) gene that is responsible for vein formation in roots, can produce proliferated bundle sheath cells as well as other changes that can be coupled to the shift to the Kranz anatomy [18]. Further work supports this relation by describing the role that the upstream interacting partner of SCR, SHORT-ROOT (SHR) plays in the variations in anatomy seen in various C₄ species [19,20].

Gene duplicates must be further refined by the mechanism by which they arise; either as single gene tandem duplication or whole genome duplication (WGD). Tandem duplications occur frequently, but the duplicates are often lost again resulting in a constant birth–death cycle of duplicate genes [21]. Second, there is whole genome duplication (WGD) or polyploidy, where all genes are simultaneously duplicated. After duplication there are often dramatic changes in the plant genomic structure, a process referred to as diploidization in which most genes return to single copy. However, the genes that are maintained in duplicate after WGD often have important functions in enzyme complexes (e.g. to maintain proper gene balance [22]) or can diversify and evolve new gene functions (e.g. neo-functionalization).

The contribution of WGD to photosynthesis-related genes has been studied in soybean, barrel-medic, Arabidopsis, and sorghum [23,24]. The polyploid and non-polyploid duplicated gene retention in Glycine max, Medicago truncatula and Arabidopsis for four classes of photosynthesis-related genes was compared: the Calvin–Benson–Bassham cycle (CBBc), the light-harvesting complex (LHC), photosystem I (PSI) and photosystem II (PSII). It was found that photosystem genes were more dosage sensitive, with more duplicates derived only from WGD whereas CC gene families were often larger with more non-polyploid duplicates retained. In Sorghum bicolor, a recent WGD was reported to be an important origin of C₄ specific genes. Several key C₄ genes of this crop were found to be collinear with genes that function in C₃ photosynthesis when compared to maize and rice. Here, we combine the approaches of these two studies to examine the evolution of photosynthesis and C₄-related genes in C₃ and C₄ Cleomaceae species.

Gynandropsis gynandra (Fig. 1, blue clade) belongs to the NAD-ME C₄ photosynthesis sub-type [25,26] and is an important South-East Asian and African dry-season leafy vegetable (sometimes referred to as Phak-sian or African cabbage), and is closely related to horticultural C₃ species Tarenaya hassleriana (Fig. 1, pink clade). Both species are easily cultivated in the greenhouse, and a robust phylogenetic framework for Cleomaceae species is emerging [4,5,27]. There are two other independent origins of the C₄ within the Cleomaceae, Cleome angustifolia and Cleome oxalidea (Fig. 1, yellow clade), identified by carbon isotope discrimination [5,25]. Because of the economic importance and ease of growth, the C₄–C₃ contrast between G. gynandra and T. hassleriana makes this system most attractive and tractable. Both species also have relatively small genome sizes (T. hassleriana ~ 292 Mb and G. gynandra ~ 1 Gb). T. hassleriana underwent a WGD named Th-α [28] but it is not yet known whether this event is shared with all or a subset of other Cleomaceae.

In this study we compare C₃ T. hassleriana of the Cleomaceae with C₄ G. gynandra of the same family. We use the knowledge of Brassicaceae gene functions to identify the important photosynthetic genes in both species and address the following questions: Does G. gynandra share the Th-α event? What is contribution of duplicate genes to photosynthesis and C₄-related gene families? And finally, what is the role of gene duplicates from WGD compared to continuous small-scale duplications?

2. Methods

2.1. Transcriptome sequencing and assembly

All transcriptome data was used directly from the Cleomaceae transcript atlas [17]. In the atlas, T. hassleriana genes were used as a reference to map transcripts from both species to Cleomaceae “unigenes” indicated by the gene name coined in the published T.
most
ences
4
defined
seq1.fasta
BlatV35
paralogs
to
–verbose
eter.

T.

strain.

TBlastX

Simplified
Fig. 1. Simplified phylogeny of Cleomaceae. Clades are numbered following the
most recently published Maximum Likelihood phylogeny of Cleomaceae [25]. Clade
15 containing T. hassleriana is marked in pink. Clade 8 containing G. gynandra is
marked in blue. Clade 5 (Yellow) contains the other origin of C4 in Cleomaceae, with
C. angustifolia and C4/C3 intermediate C. paradoxa. (For interpretation of the references
to color in this figure legend, the reader is referred to the web version of the article.)

hassleriana

2.2. Homolog selection

A TBlastX [31] search of transcriptomes of T. hassleriana and G.
gynandra was performed with default parameters (no e-value cutoff) to have a maximum number of hits for subsequent filtering. To filter paralogs and orthologs from these results, CIP/CALP filtering was used [32]. Cumulative Identity Percentage (CIP) is defined as the sum of the number of matching nucleotides for each high-scoring segment pair (HSP) of a pair of genes divided by the total lengths of those HSPs. Cumulative alignment length percentage (CALP) is defined as the sum of the alignment lengths of all HSPs of a matching gene pair divided by the total length of the query sequence. Both of these values give a reliable estimation of the similarity of two genes and is a more accurate method than evalue or bit score threshold filtering. A CIP/CALP threshold of 50/50 was chosen as a suitable cutoff point for orthology and/or paralogy.

2.3. Ks/4dvt calculation of paralog pairs

Paralogs identified with CIP/CALP filtering were aligned using Exonerate [33] with the coding2coding model parameter, using a custom output format through the “roll your own” parameter. The exact command line used was: “exonerate -m c2c seq1.fasta seq2.fasta -ryo \"%Pqs %Pts \n\" -showalignment false –verbose 0”. The output from this command was fed into CodeML from the PAML package using standard parameters (Codonfreq = 2, kappa = 2, omega = 0.4). Output from PAML [34] was parsed using custom Perl scripts to read the synonymous substitution rate (Ks) and the fourfold transition rate (4dvt). This workflow is identical to the established paralog identification pipeline Duppipe [35] using updated tools and more stringent selection using CIP/CALP.

2.4. Homolog clustering

Photosynthesis genes were selected from known functionally
annotated Arabidopsis genes. Gene identifiers for each family are listed hereafter and in Table 2. BCA: AT1G23730, AT1G58180, AT1G70410, AT3G01500, AT4G35580, AT5G14740. MDH (cytosolic): AT1G04410, AT5G43330, AT5G56720. MDH (mitochondrial): AT1G53240, AT2G22780, AT3G15020, AT3G47520, AT5G09660. MDH (plastidic): AT1G53240, AT2G22780, AT3G15020, AT3G47520, AT5G09660. NAD-ME: AT2G13560, AT4G00570. NADP-ME: AT1G79750, AT2G19900, AT5G11670, AT5G25880. PEPC: AT1G21440, AT1G53310, AT2G42600, AT3G14940. PPCK: AT1G08650, AT3G04530, AT3G04550, AT4G37870, AT5G28500, AT5G65690. These genes were then used as a BLAST database and queried with T. hassleriana and G. gynandra atlas unigenes. Hits were then filtered using a 50/50 CIP/CALP cutoff. Using custom Perl scripts, the hits of these hits were picked up, iterating recursively until convergence (no new hits found). All unique genes resulting from this process form a family cluster.

2.5. Synteny analyses

T. hassleriana genes were used as a query in the CoGe Synfind
[36] program using the following parameters: Comparison algo-
rithm: Last. Gene window size: 40. Minimum number of genes: 4. Scoring Function: Collinear, Syntenic depth: unlimited. As query genomes, the following were used: A. arabicum VEGI unmasked v2.5, A. thaliana Col-0 TAIR unmasked v10.02 and T. hassleriana BGI; Eric Scranz Lab; Weber lab unmasked v5.

3. Results

3.1. Evidence of WGD in both species confirming a shared event

Using the transcript sets of G. gynandra and T. hassleriana, paralogs were matched to each other by BLAST search and CIP/CALP filtering. In total, 55,014 paralogs were found: 26,883 in T. hassleriana covering 49% of transcript space and 28,131 in G. gynandra covering 48% of transcript space. Of all paralog pairs, Ks and fourfold transversion substitutions (4dvt) were determined and binned to establish an evolutionary time distribution (Fig. 2). In both species a large gene birth event has taken place around Ks = 0.4 (Fig. 2 between Ks = 0.25 and Ks = 0.5), which corresponds to the Ks window established earlier for the Th-α paleohaploidy event [28]. The same analysis was performed using 4dvt values and results were extremely similar. Enumerating the paralogs that fall within the Th-α peak, we see that 15,785 gene pairs in T. hassleriana are retained from the Th-α paleohaploidy, or ~29% of the total transcriptome. For G. gynandra, 16,096 gene pairs fall within the Th-α window, or around 27% of all transcripts.

3.2. Duplicate loss and retention in essential C4 families

We examined six gene families that are essential in C4 photosynthesis in detail: NAD malic enzyme (NAD-ME), NADP malic enzyme (NADP-ME), β carbonic anhydrase (B3A), malate dehydrogenase (MDH), phosphoenolpyruvate carboxylase (PEPC) and phosphoenolpyruvate carboxykinase (PPCK). Using Arabidopsis genes as a reference, homologous clusters were created using a
CIP/CALP cutoff of 50/50. 146 homologous pairs could be placed in a cluster across the three species comprising 105 unique genes (Table 1); 40 in A. thaliana, 57 in T. hassleriana and 49 in G. gynandra. In most cases both Cleomaceae species have around 1.5 times the number of genes of A. thaliana except, interestingly, the NADP-ME family where numbers are almost the same in all species. Also of note is that T. hassleriana has 16% more C4 related genes in total than G. gynandra (57 over 49).

All genes of one species in a cluster were then aligned to each other and the Ks value of each pairing was established and subsequently binned with a stepsize of $Ks = 0.15$ (Fig. 3). At the $Ks$ corresponding to the Th-α hexaploidy, both T. hassleriana and G. gynandra show a relative increase of gene pairs with this amount of synonymous substitutions. A. thaliana at the $Ks$ of its older At-α event shows a similar, if slightly lower increase. Even longer ago in

![Fig. 2](https://example.com/image2.png)

**Fig. 2.** Histogram showing the amount of gene pairs per $Ks$ bin for T. hassleriana (pink) and G. gynandra (blue). The peak at around $Ks = 0.45$ is an indication of a massive gene birth event and is considered evidence of paleopolyploidy. Both species have an extremely similar peak, indicating that this is a shared polyploidy event. The $Ks$ values of these peaks correspond with $Ks$ values found earlier for the Th-α hexaploidy event, indicating that this event has occurred before divergence of T. hassleriana and G. gynandra. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

![Table 1](https://example.com/table1.png)

**Table 1**

|        | A. thaliana | T. Hassleriana | G. gynandra |
|--------|-------------|----------------|-------------|
| βCA    | 6           | 10             | 7           |
| MDH (cyt.) | 3        | 6              | 6           |
| MDH (mit.) | 5        | 6              | 6           |
| MDH (per.) | 5        | 8              | 6           |
| MDH (plast.) | 5      | 6              | 6           |
| NAD-ME | 2           | 3              | 3           |
| NADP-ME | 4           | 4              | 3           |
| PEPC   | 4           | 8              | 6           |
| PPCK   | 6           | 6              | 6           |
| **Total** | **40**     | **57**         | **49**      |

C4 photosynthesis homolog cluster sizes in A. thaliana, T. hassleriana and G. gynandra. Both Cleomaceae species have around 1.5 times the number of genes of A. thaliana except the NADP-ME and NAD-ME families where numbers are lower than average in the Cleomaceae species resulting in a similar amount of homologs in each species for these two gene groups.

![Fig. 3](https://example.com/image3.png)

**Fig. 3.** Histogram showing Ks values of homolog gene clusters associated with C4 photosynthesis: MDH, NAD-ME, NADP-ME, PEPC and βCA. Gene duplication events are marked at their associated Ks value and colored according to earlier publication [28]; a square indicates a duplication (tetraploidy), a circle indicates a triplication (hexaploidy). The contribution of the Th-α (pink circle) and the At-α (orange square) on photosynthesis related gene copy number can be seen at $Ks = 0.45$ and $Ks = 0.6$ respectively. The β event at $Ks = 1.8$ (blue square) has contributed substantially to the expansion of gene copy number in T. hassleriana. Further in evolutionary time, around $Ks = 2.4$, the γ event (green circle) that is also shared by all three species has contributed equally to the polyploid presence in photosynthetic orthologs. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)
evolutionary time at the Ks corresponding to the β event T. hasselliana has retained ~20% of C₄ related genes, where the other species show 2% and 0% retention for G. gynandra and A. thaliana, respectively. The final confirmed paleohexaploidy that all three species share, the ancient γ event at Ks = 2.4, has contributed substantially to the genetic makeup of all three species. In A. thaliana the number of relations that stem from the γ paleohexaploidy is 23%, with both Cleomaceae at 15% and 21% for T. hasselliana and G. gynandra, respectively.

3.3. Syntenic copy number variation

Syntenic analyses of the previously mentioned gene families was performed using CoGe Synfind [36]. Each T. hasselliana C₄ related ortholog was used as a query with T. hasselliana, Arabidopsis thaliana, A. arabicum [37] as a basal representative of Brassicaceae. Thus for the T. hasselliana: A. thaliana: A. arabicum ortholog ratio we would theoretically expect 3 (Th-α):2 (At-α):2. Query results were enumerated and the average number of regions per family was determined (Fig. 4). For many families, the average is comparable to the 3:2:2 ratio, which is also represented by the average ratio (Fig. 4, rightmost set of bars) being 3.6:2.1:2.5. The exception is NAD-ME family, which has seen more than expected retention with an orthologs ratio 4.3:3:3:4.3. The PEPC family also seems slightly under-retained in Brassicaceae, with a ratio of 3:1:3:1:6. Unfortunately, syntenic data is impossible to obtain without a sequenced genome so data syntenic regions of G. gynandra will have to be obtained in future work.

3.4. Regulation of photosynthetic homolog expression

Both Cleomaceae have substantially more copies of photosynthetic genes (Fig. 4). Using the Cleomaceae expression atlases [17], the expression of separate copies was compared in the C₃ and the C₄ species. In the expression atlas, the T. hasselliana coding sequence was used as a reference to map expression in both T. hasselliana and G. gynandra to a single Cleomaceae ‘unigene’. Expression was quantified in nine different tissues including three developmental series: development from young to mature leaf (six stages), root, stem, stamen, petal, carpel, sepal, a seedling developmental series (three stages) and a seed time series (three stages).

For the photosynthetic gene families (NAD-ME, NADP-ME, PEPCK, PEPC, MDH, CA), homolog selection resulted in a data set of 43 unigenes with expression data for both Cleomaceae species. Expression levels were normalized and compared amongst photosynthetic gene families, examples of which are plotted for NAD-ME and βCA (Fig. 5). Immediately noticeable is the highly similar expression profiles of G. gynandra when compared to the more chaotic profiles of T. hasselliana. This is observed in all except one gene family, G. gynandra has 176 expressed unigenes with a highly correlated expression pattern (Pearson correlation > 0.95) whereas in T. hasselliana 87 unigenes share a highly correlated expression pattern (Pearson correlation > 0.95).

The expression pattern that is observed in G. gynandra in the β-CA family also correspond to their A. thaliana highest ranking match (Table 2). The cluster consisting of C.spinososa_00253, C.spinososa_13896, C.spinososa_18526 and C.spinososa_10164 for example all match highest to A. thaliana gene β carbonic anhydrase 4 (AT1G70410). The cluster consisting of C.spinososa_07642 and C.spinososa_13410 both map to carbonic anhydrase 1 (AT3G01500). A similar pattern is present in NAD-ME where the cluster of C.spinososa_03046 and C.spinososa_09126 both map to NAD-ME1 (AT2G13560) and the C.spinososa_12536 singleton maps to NAD-ME2 (AT4G00570).

### Table 2

| Gene family | ATG identifiers |
|-------------|-----------------|
| BCA         | AT1G23730       |
|             | AT1G58180       |
|             | AT1G70410       |
|             | AT3G01500       |
|             | AT4G33580       |
|             | AT5G14740       |
| MDH (cytosolic) | AT1G04410   |
|             | AT5G14740       |
|             | AT5G56620       |
| MDH (mitochondrial) | AT1G53240 |
|             | AT2G22780       |
|             | AT3G15020       |
|             | AT3G47520       |
|             | AT5G09660       |
| MDH (plastidic) | AT1G53240   |
|             | AT2G22780       |
|             | AT3G15020       |
|             | AT3G47520       |
|             | AT5G09660       |
| NAD-ME       | AT2G13560       |
|             | AT4G00570       |
| NADP-ME      | AT1G79750       |
|             | AT2G19900       |
|             | AT5G11670       |
|             | AT5G25880       |
| PEPC         | AT1G21440       |
|             | AT1G53310       |
|             | AT2G42600       |
|             | AT3G14940       |
| PPCK         | AT1G08650       |
|             | AT3G04530       |
|             | AT3G04550       |
|             | AT4G37870       |
|             | AT5G28500       |
|             | AT5G06590       |

![Fig. 4](image-url) Histogram showing average syntenic region copy number for T. hasselliana, A. thaliana and Aethionema arabicum. Because A. arabicum and A. thaliana both share a paleotetraploidy, the expected ratio of syntenic regions for T. hasselliana: A. thaliana: Aethionema arabicum is 3:2:2. In most cases, syntenic regions follow this distribution which is also reflected in the average ratio of all families being 3.6:2.1:2.5 (rightmost bars). The exception is NAD-ME, where the average region number in both A. arabicum and A. thaliana is as high as T. hasselliana.
Fig. 5. Canalization in expression of NAD malic enzyme (top and bottom left) and \( \beta \) carbonic anhydrase (top and bottom right) homologs in *T. hassleriana* and *G. gynandra*. Top left: NAD-ME expression in *T. hassleriana*. Top right: \( \beta \)CA expression in *T. hassleriana*. Bottom left: NAD-ME expression in *G. gynandra*. Bottom right: \( \beta \)CA expression in *G. gynandra*. (Mapped) gene names and associated colors are displayed, see Materials and Methods for more details on the mapping of *G. gynandra* transcripts to *T. hassleriana* genes. Note that leaf0–leaf5 as well as seedling2–seedling6 and seed1–seed3 are time series of the same organ, with the leaf and seedling gradient being two days separated by stage. Transcription levels in *G. gynandra* (lower graphs) are more strictly regulated across organs, seeds and seedlings. The chaotic patterns in *T. hassleriana* (upper graphs) results in half the genes having a Pearson correlation >0.95 compared to *G. gynandra*.

### 4. Discussion and conclusions

In this study, we have analyzed the transcriptomes of the C\(_3\) *T. hassleriana* and C\(_4\) *G. gynandra* to address the potential contribution of WGD and recent gene duplicates to the evolution of photosynthesis and C\(_4\)-pathway related genes. The initial comparison of *T. hassleriana* and *G. gynandra* was performed to identify the differential expression of key-genes involved in the NAD-ME C\(_4\) biochemical pathway. However, it did not consider the role of gene duplicates. We show that very distinct patterns will occur when the duplication history is taken into account.

We could confirm the Th-\( \alpha \) hexaploidy that has been found in *T. hassleriana* using an independent transcriptome dataset. We also find that *G. gynandra* shares this WGD with *T. hassleriana*, further establishing the occurrence of WGD in this lineage. Based on the phylogenetic position of both species in Cleomaceae, the Th-\( \alpha \) duplication took place at least before the divergence of the two species which means that it is shared across Cleomaceae lineages 8–15 according to the latest phylogeny of the family [25]. Dating this polyploidy event in terms of absolute age is always a difficult task, however, here we find that the KS rate of *G. gynandra* is extremely similar if not identical to *T. hassleriana*. Assuming then that mutation rates between these two species are the same, we can reconfirm the previous date estimation of Th-\( \alpha \) at 13.7 mya [38].

The influence of the Th-\( \alpha \) WGD event on photosynthetic gene composition is apparent, both in ortholog number as well as in syntenic region copy number for both species. From absolute orthologs numbers we can see that there is no increased retention between Cleomaceae species and even a slightly lower rate of retention in *G. gynandra*. This indicates that both species have experienced similar evolutionary constraints for a significant amount of time. Also we need to consider that genes sharing a similar sequence, do not necessarily have to share the same function. Even using strict CIP/CALP filtering which has been proved to be an accurate measure for the prediction of true orthologs [32], differential expression either in time, localization or regulation can substantially change the function of a gene. This is especially the case for genes in the core C\(_4\) photosynthesis pathway, where many C\(_3\) genes have been recruited into new functions [13,39].

When establishing KS values of deeper ortholog nodes of photosynthesis genes, a large proportion of genes seem to have been retained from the \( \gamma \) duplication. For a trait that is likely to be highly dosage sensitive [23], we expect that gene loss will be rare and that remnants from this old paleohexaploidy are still present. However, considering the time that has passed since the \( \gamma \) paleohexaploidy event and on the basis of absolute gene copy numbers some gene loss has taken place predating the transition from C\(_3\) to C\(_4\).

The evolutionary importance of WGD events is made clear from the dominant presence of retained Th-\( \alpha \) genes in both Cleomaceae species. However, certain questions remain: Can we couple this importance to the evolution of specific traits or in this case, C\(_4\) photosynthesis? This is an old discussion, dating back to the works of
Ohno who was the first to suggest that the massive radiation of vertebrates was caused by a whole genome duplication in the ancestor [40]. An earlier study on the evolution of photosynthesis in soybean, showed that the Calvin–Benson–Bassham cycle (CBB) and the light harvesting complex (LHC) gene families show a greater yield from single gene duplications than both photosystem groups. This is explained by the increased dosage sensitivity of photosystem genes: if some subunits are expressed differently due to duplications while others are not, this is deleterious for the system as a whole [23]. This acts as a conservation mechanism for gene copy number that does not affect the more loosely connected enzyme collection of the CBB and LHC genes.

In C. gynandra, where the expression of C4 genes is tightly linked in clusters we would expect a high retention of orthologs. However, this dependency on transcriptional regulation has not lead to an increased retention of photosynthetic genes, as evidenced by lower copy numbers for all C4 gene families when compared to T. hassleriana. It is not likely that neofunctionalization of genes after polyploidy has played a major role in the shift to C4 photosynthesis. The much more stringent transcriptional regulation of C4 cycle genes in C. gynandra when compared to T. hassleriana as evidenced in this study is in accordance with the alternative hypothesis, which states that this process was mainly due to recruitment of existing genes in transcriptional space as suggested by several authors [12,14,41,42].

We still have much to learn regarding the development of C4 photosynthesis. When studying this exceptional trait, we must always consider the genetic history of the species in question. Here, we give evidence that duplications, on a large scale and small, contribute to trait evolution. The exact mechanisms behind the recruitment of these genes into new biochemical pathways however are still largely unknown. Current sequencing efforts for C. gynandra will significantly aid in finding the detailed mechanisms of gene and C4 photosynthesis evolution. The Cleome genus provides an excellent model system for unraveling the evolutionary origin and workings of C4 photosynthesis and hopefully will enable us to harvest the fruits of our knowledge on this remarkable form of plant energy conversion.

Competing interests

The authors declare that they have no competing interests.

Authors’ contributions

Cleome transcriptome sequencing, processing, assembly and quantification was done by CK, AB and APMW provided comments on handling highly expressed duplicates as well as proofreading the manuscript. EvdB performed the bioinformatic analyses, EvdB and MES prepared the manuscript. JMH and ZX proofread and edited the manuscript.

Acknowledgements

The work of EvB and MES was funded by NWO Vernieuwingssimpuls Vidi grant number 864.10.001. APMW acknowledges support by the Deutsche Forschungsgemeinschaft (SPP 1529: EXC 1028).

References

[1] R.F. Sage, P.-A. Christin, E.J. Edwards, The C4 plant lineages of planet Earth, J. Exp. Bot. 62 (2011) 3155–3169.
[2] R.F. Sage, The evolution of C4 photosynthesis, N. Phytol. 161 (2004) 341–370.
[3] M.S. Ku, J. Wu, Z. Dai, R.A. Scott, C. Chu, G.E. Edwards, Photosynthetic and photosynthetic characteristics of Flaveria species, Plant Physiol. 96 (1991) 518–528.
[4] N.J. Brown, K. Parsley, J.M. Hibberd, The future of C4 research – maize, Flaveria and Cleome? Trends Plant Sci. 10 (2005) 215–221.
[5] D.M. Marshall, R. Muhaidat, N.J. Brown, Z. Liu, S. Stanley, H. Griffiths, R.F. Sage, J.M. Hibberd, Cleome: a genus closely related to Arabidopsis, contains species spanning a developmental progression from C3 to C4 photosynthesis, Plant J. 51 (2007) 886–896.
[6] X.-G. Zhu, S.P. Long, D.R. Ort, Improving photosynthetic efficiency for greater yield, Annu. Rev. Plant Biol. 61 (2010) 235–261.
[7] G.E. Edwards, V.R. Franceschi, E.V. Vozнесенская, Single-cell C4 photosynthesis versus the dual-cell (Kranz) paradigm, Annu. Rev. Plant Biol. 52 (2004) 173–196.
[8] A.D. McKown, N.G. Dengler, Vein patterning and evolution in C4 plants, Botany 88 (2010) 775–786.
[9] Y. Wang, A. Bräutigam, A.P.M. Weber, X.-G. Zhu, Three distinct biochemical subtypes of C4 photosynthesis? A modelling analysis, J. Exp. Bot. 65 (2014) 3567–3578.
[10] J.R. Ehleringer, T.E. Cerling, B.R. Helliker, C4 photosynthesis, atmospheric CO2, and climate, Oecologia 112 (1997) 285–299.
[11] A. Bräutigam, K. Kajala, J. Wullenweber, M. Sommer, D. Gagneul, K.L. Weber, K.M. C4 photosynthesis derived from comparative transcriptomics of closely related C3 and C4 species, Plant Physiol. 155 (2011) 142–156.
[12] N.J. Brown, C.A. Newell, S. Stanley, J.E. Chen, A.J. Perrin, K. Kajala, J.M. Hibberd, Independent and parallel recruitment of preexisting mechanisms underlying C4 photosynthesis, Science 331 (2011) 1436–1439.
[13] J.M. Hibberd, S. Covshoff, The regulation of gene expression required for C4 photosynthesis, Annu. Rev. Plant Biol. 61 (2010) 181–207.
[14] K. Kajala, N.J. Brown, B.P. Williams, P. Borrill, L.E. Taylor, J.M. Hibberd, Multiple Arabidopsis genes primed for recruitment into C4 photosynthesis, Plant J. 69 (2012) 47–56.
[15] K. Monson Russell, Gene duplication, neofunctionalization, and the evolution of C4 photosynthesis, Int. J. Plant Sci. 164 (2003) 543–554.
[16] R.K.S. Monson, F. Rowan, The origins of C4 genes and evolutionary pattern in the C4 metabolic phenotype, in: Academic Press (Ed.), C4 Plant Biology, 1999, pp. 377–410, San Diego.
[17] C. Külhaboglu, A.K. Denton, M. Sommer, J. Maal, S. Schliesky, T.J. Wrobel, B. Berckmans, E. Gongora-Castillo, C.R. Buel, R. Simon, et al., Comparative Transcriptome Atlases Reveal Altered Gene Expression Modules Between Two Cleome Species C3 and C4 Plants, The Plant Cell Online (2014), http://dx.doi.org/10.1105/tpc.114.123752, Advance online publication.
[18] T.L. Slewinski, A.A. Anderson, C. Zhang, R. Turgeon, Scarcecrow plays a role in establishing Kranz anatomy in maize leaves, Plant Cell Physiol. 53 (2012) 2693–2697.
[19] T.L. Slewinski, A.A. Anderson, S. Price, J.R. Withee, K. Gallagher, R. Turgeon, Short-root1 plays a role in the development of vascular tissue and Kranz anatomy in maize leaves, Mol. Plant 7 (2014) 1388–1392.
[20] P. Wang, S. Kelly, J.P. Pourcare, J.A. Langdale, Genome-wide transcript analysis of early maize leaf development reveals gene cohorts associated with the differentiation of C4 kranz anatomy, Plant J. 75 (2013) 656–670.
[21] S.B. Cannon, A. Mitra, A. Baungarten, N.D. Young, M. Gay, The roles of segmental and tandem gene duplication in the evolution of large gene families in Arabidopsis thaliana, BMC Plant Biol. 4 (2004) 10.
[22] P. Edger, J.C. Pires, Gene and genome duplications: the impact of dosage compensation on the fate of nuclear genes, Chromosome Res. 17 (2009) 699–717.
[23] J.E. Coate, J.A. Schlueter, A.M. Whaley, J.J. Doyle, Comparative evolution of photosynthetic genes in response to polyploid and nonpolyploid duplication, Plant Physiol. 155 (2011) 2081–2095.
[24] X. Wang, U. Gowik, H. Tang, J.E. Bowers, P. Westhoff, A.H. Paterson, Comparative genomic analysis of C4 photosynthetic pathway evolution in grasses, Genome Biol. 10 (2009) 658.
[25] T. Fedorov, E.V. Vozнесенская, G.E. Edwards, E.H. Ra了好ison, Biogeographic patterns of diversification and the origins of C4 in Cleome (Cleomaceae), Syst. Bot. 35 (2010) 811–826.
[26] E.V. Vozнесенская, N.K. Koteyeva, S.D. Choung, A.N. Ivanova, J. Barroca, L.A. Craven, G.E. Edwards, Physiological, anatomical and biochemical characterisation of photosynthetic types in genus Cleome (Cleomaceae), Funct. Plant Biol. 34 (2007) 247–267.
[27] R.D. Marquard, R. Steinback, A model plant for a biology curriculum: spider flower (Cleonex hassleriana L.), Am. Biol. Teach. 71 (2009) 235–244.
[28] M.S. Barker, H. Vogel, M.E. Schranz, Paleopolyploidy in the Brassicaceae: analyses of the cleome transcriptome elucidate the history of genome duplications in Arabidopsis and other Brassicaceae, Genome Biol. Evol. 1 (2009) 391–399.
[29] S. Cheng, E. van den Bergh, P. Zeng, X. Zhong, J. Xu, X. Liu, J. Hoffberger, S. de Bruijn, A.S. Bhide, C. Kueelahoglu, The Tarenaya hasslerianna genome provides insight into reproductive trait and genome evolution of crucifers, Plant Cell Online 25 (2013) 2813–2830.
[30] W.J. Kent, BLAT – the BLAST-like alignment tool, Genome Res. 12 (2002) 656–664.
[31] J.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W.M. Miller, D.J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res. 25 (1997) 3389–3402.
[32] F. Murat, Y. Van de Peer, J. Salse, Decoding plant and animal genome plasticity from differential paleo-evolutionary patterns and processes, Genome Biol. Evol. 4 (2012) 917–928.
[33] G.S. Slater, E. Birney, Automated generation of heuristics for biological sequence comparison, BMC Bioinform. 6 (2005) 31.
Z. Yang, PAML: a program package for phylogenetic analysis by maximum likelihood, Comput. Appl. Biosci. 13 (1997) 555–556.

M.S. Barker, K.M. Dlugosch, L. Dinh, R.S. Challa, N.C. Kane, M.G. King, L.H. Rieseberg, EvoPipes.net. Bioinformatic tools for ecological and evolutionary genomics, Evol. Bioinform. 6 (2010) 143–149.

E. Lyons, M. Freeling, How to usefully compare homologous plant genes and chromosomes as DNA sequences, Plant J. 53 (2008) 661–673.

A. Haudry, A.E. Platts, E. Vello, D.R. Hoen, M. Leclercq, R.J. Williamson, E. Forczek, Z. Joly-Lopez, J.G. Steffen, K.M. Hazzouri, et al., An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions, Nat. Genet. 45 (2013) 891–898 (advance online publication).

M.S. Barker, H. Vogel, M.E. Schranz, Paleopolyploidy in the Brassicales: analyses of the Cleome transcriptome elucidate the history of genome duplications in Arabidopsis and other Brassicales, Genome Biol. Evol. 1 (2009) 391.

U. Gowik, J. Burscheidt, M. Akyildiz, U. Schlué, M. Kocor, M. Streubel, P. Westhoff, cis-Regulatory elements for mesophyll-specific gene expression in the C4 plant Flaveria trinervia, the promoter of the C4 phosphoenolpyruvate carboxylase gene, Plant Cell Online 16 (2004) 1077–1090.

S. Ohno, U. Wolf, N.B. Atkin, Evolution from fish to mammals by gene duplication, Hereditas 59 (1968) 169–187.

U. Gowik, P. Westhoff, The path from C3 to C4 photosynthesis, Plant Physiol. 155 (2011) 56–63.

B.P. Williams, S. Aubry, J.M. Hibberd, Molecular evolution of genes recruited into C4 photosynthesis, Trends Plant Sci. 17 (2012) 213–220.

P. Lamesch, T.Z. Berardini, D. Li, D. Swarbreck, C. Wilks, R. Sasidharan, R. Muller, K. Dreher, D.L. Alexander, M. Garcia-Hernandez, et al., The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools, Nucleic Acids Res. 36 (2011) D1009–D1014.