Current Plant Biology 1 (2014) 2–9

Contents lists available at ScienceDirect

Current Plant Biology

journal homepage: www.elsevier.com/locate/cpb

Gene and genome duplications and the origin of C_4 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, Droevendaalsesteeg 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
C_4 photosynthesis

ABSTRACT

C_4 photosynthesis is a trait that has evolved in 66 independent plant lineages and increases the efficiency of carbon fixation. The shift from C_3 to C_4 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 C_4 photosynthesis, with its C_3 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 C_3 and C_4 species exists in terms of gene copy number. This is further confirmed by syntenic 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 C_4 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 C_4 photosynthesis is not significantly influenced by polyploidy combined with the fact that transcript abundance in *G. 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 C_4 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 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

1. Introduction

Over sixty lineages of both monocot and eudicot angiosperms have evolved a remarkable solution to maximize photosynthesis efficiency under low CO_2 levels, high temperatures and/or drought: C_4 photosynthesis [1]. The evolution of this modified photosynthetic pathway represents a wonderful example of convergent evolution. While the changes necessary for the transition from C_3 to C_4 photosynthesis are numerous, the trait has a wide phylogenetic distribution across angiosperms, with 19 different plant

* Corresponding author. Tel.: +31 317483160.
E-mail addresses: erik.vandenbergh@wur.nl (E. van den Bergh), canan.kuelahoglu@uni-duesseldorf.de (C. Külahoglu), andrea.braeutigam@uni-duesseldorf.de (A. Bräutigam), jmh65@cam.ac.uk (J.M. Hibberd), andreas.weber@uni-duesseldorf.de (A.P.M. Weber), zhuxinguang@picb.ac.cn (X.-G. Zhu), eric.schranz@wur.nl (M. Eric Schranz).

http://dx.doi.org/10.1016/j.cpb.2014.08.001

families across the globe known to contain one or multiple members capable of C_4 photosynthesis [2]. Much research on eudicot C_4 has focused on *Flaveria* species (Asteraceae), which contains not only C_4 species but also a number of C_3/C_4 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 C_4 [4,5].

C_4 plants spatially separate the fixation of carbon away from the RuBisCO active site by using phospho*enol*pyruvate carboxylase, an alternate carboxylase that does not react with oxygen. As a consequence they are more efficient under permissive conditions [6]. The typical C_4 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 C_4 leaves is increased [8]. This internal leaf architecture physically partitions the biochemical events of the C_4 pathway into two main phases. In the first phase, dissolved HCO_3^- is assimilated into C_4 acids by phospho*enol*pyruvate 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 CO_2 is fixed by RuBisCO. The increased CO_2 concentration in the BS cells allows carbon fixation by RuBisCO to be much more efficient by reducing photorespiration. Two subtypes of the C_4 biochemical pathway are defined, based on the most active C_4 acid decarboxylase that liberates CO_2 from C_4 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 C_4.

The process of carboxylation and decarboxylation costs more energy than the simpler C_3 form of photosynthesis, but it diminishes photorespiration. In conditions of low atmospheric CO_2 pressure, photorespiration causes a major loss in photosynthetic output and the elaborate concentrating mechanisms of C_4 photosynthesis circumvent this [10].

All genes important for the C_4 pathway are expressed at relatively low levels in C_3 leaves [11]. The mechanism for recruitment of these genes into the C_4 pathway remains to be elucidated. For some ancestral C_3 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 C_4 pathway. It has been proposed that gene duplication and subsequent neofunctionalization of one gene copy has facilitated the alterations in gene expression that underlie the evolution of C_4 photosynthesis [15,16]. Gene duplication is proposed to be a (pre)condition for the evolution of C_4 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 C_4 biochemistry, but rather changes in expression and localization of specific genes [11,17]. However, this study highlighted just the number of C_4 genes and did not take into account the age and mechanism of gene duplications.

The modifications necessary for the anatomical changes from C_3 to C_4 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 C_4 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 C_4 specific genes. Several key C_4 genes of this crop were found to be collinear with genes that function in C_3 photosynthesis when compared to maize and rice. Here, we combine the approaches of these two studies to examine the evolution of photosynthesis and C_4-related genes in C_3 and C_4 Cleomaceae species.

Gynandropsis gynandra (Fig. 1, blue clade) belongs to the NAD-ME C_4 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 C_3 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 C_4 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 C_4–C_3 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 C_3 *T. hassleriana* of the Cleomaceae with C_4 *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 C_4-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.*