Evolution of C_3-C_4 intermediate photosynthesis in Brassicaceae

Inaugural-Dissertation

zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Heinrich-Heine-Universität Düsseldorf

vorgelegt von

Sebastian Triesch
aus Solingen

Düsseldorf, Juli 2024

aus dem Institut für Biochemie der Pflanzen
der Heinrich-Heine-Universität Düsseldorf

Gedruckt mit der Genehmigung der
Mathematisch-Naturwissenschaftlichen Fakultät der
Heinrich-Heine-Universität Düsseldorf

Berichterstatter:

1. Prof. Dr. Andreas P. M. Weber

2. Prof. Dr. Stanislav Kopriva

Tag der mündlichen Prüfung: 17.12.2024

Eidesstattliche Erklärung

Ich versichere an Eides Statt, dass die Dissertation von mir selbständig und ohne unzulässige fremde Hilfe unter Beachtung der „Grundsätze zur Sicherung guter wissenschaftlicher Praxis an der Heinrich-Heine-Universität Düsseldorf“ erstellt worden ist.

Die Dissertation habe ich in der vorgelegten oder in ähnlicher Form noch bei keiner anderen Institution eingereicht.

Ich habe bisher keine erfolglosen Promotionsversuche unternommen.

Düsseldorf, den 31.07.2024

Sebastian Triesch

1

Summary

Plant species with C_3-C_4 intermediate photosynthesis can be seen as stable interim stages on the evolutionary path towards the efficient C_4 photosynthesis. These species show an increased number of plastids in bundle sheath cells that are organized around the leaf veins in a Kranz-like anatomy. Moreover, C_3-C_4 intermediate species can utilize the glycine shuttle between the mesophyll and bundle sheath cells, a carbon concentrating mechanism that proportionally decreases the oxygenation reaction of Rubisco and thereby reduces flux through the costly photorespiratory pathway. Like all complex traits, C_3-C_4 intermediate photosynthesis relies on the spatially differentiated expression of genes, especially between mesophyll and bundle sheath cells. With the goal to unravel the genetic background of C_3-C_4 intermediate photosynthesis, we analyzed the cell-specific gene expression and its underlying functional genetics in the Brassicaceae family that contains model species as well as crop plants. In the *bona fide* first single-nuclei transcriptome study for a C_3-C_4 intermediate plant, we found a significant recruitment of gene expression from mesophyll to bundle sheath cells. Especially the expression of genes involved in photorespiratory glycine decarboxylation was shifted to bundle sheath cells. Using a pan-genomic association study, we correlated this expression shift to the insertion of transposable elements in the upstream region of the *GLDP1* gene. This insertion is a large convergent evolutionary event in a polyphyletic clade of species with C_3-C_4 intermediacy. We showed *in vivo* that the transposon integration contributes mechanistically to the *GLDP1* expression shift to the bundle-sheath cells, presumably in coordination with other unknown genetic and epigenetic factors. Our results shed light on the mechanistic of the early steps in the evolution of C_3-C_4 intermediate photosynthesis in the Brassicaceae family, especially on the genetics of bundle sheath cell specificity.

2 Zusammenfassung

Pflanzenarten mit C_3-C_4 intermediärer Photosynthese können als stabile Zwischenstufen auf dem evolutionären Weg zur effizienten C_4-Photosynthese betrachtet werden. Diese Arten zeigen eine erhöhte Anzahl von Plastiden in Bündelscheidenzellen, die um die Blattadern in einer Art Kranz-Anatomie angeordnet sind. Darüber hinaus können C_3-C_4 intermediäre Arten einen Glycin-Shuttle zwischen Mesophyll- und Bündelscheidenzellen nutzen, einen Kohlenstoffkonzentrationsmechanismus, der die Reaktion von Rubisco mit Sauerstoff proportional verringert und somit die energieaufwändige Photorespiration reduziert. Die C_3-C_4 intermediäre Photosynthese beruht auf räumlich differenzierter Genexpression, insbesondere zwischen Mesophyll- und Bündelscheidenzellen. Mit dem Ziel, die genetischen Grundlagen der C_3-C_4 intermediären Photosynthese zu entschlüsseln, analysierten wir die zellspezifische Genexpression und deren zugrunde liegende funktionelle Genetik in der Familie der Brassicaceae, die sowohl Modellarten als auch Nutzpflanzen umfasst. In der unseres Wissens nach ersten Einzelzelltranskriptom-Studie einer C_3-C_4 intermediären Pflanze fanden wir eine signifikante Rekrutierung der Genexpression von Mesophyll- zu Bündelscheidenzellen. Insbesondere die Expression von Genen, die an der photorespiratorischen Glycindecarboxylierung beteiligt sind, verlagerte sich in die Bündelscheidenzellen. Mit Hilfe einer pan-genomischen Assoziationsstudie wurde diese Expressionsverschiebung mit der Insertion von Transposons in den *upstream*-Bereich des *GLDP1*-Gens korelliert. Diese Insertion stellt ein signifikantes konvergentes evolutionäres Ereignis in einer polyphyletischen Klade von Arten mit C_3-C_4-intermediärer Photosynthese dar. Wir zeigten *in vivo*, dass die Transposon-Integration mechanistisch zur *GLDP1*-Expressionsverschiebung in die Bündelscheidenzellen beiträgt, vermutlich in Koordination mit anderen unbekannten genetischen und epigenetischen Faktoren. Unsere Ergebnisse werfen Licht auf die Mechanismen der frühen Schritte in der

Evolution der C_3-C_4 intermediären Photosynthese in der Familie der Brassicaceae, insbesondere auf die Genetik der Bündelscheidenzellspezifität.

3 Acknowledgements

It is incredible how many people I had the chance to meet, collaborate with and trust over the last four years. I want to take this little chapter to say thanks to a few of them.

First of all, I want to thank my thesis advisor **Prof. Dr. Andreas P. M. Weber** for his support, guidance and trust. I highly value that I had the opportunity to shape my project according to my wishes, had the chance to run large-scale sequencing experiments and visit conferences all around Europe.

Special thanks go to **Dr. Urte Schlüter** for being my supervisor on daily basis. Thank you for guiding me through the complex world of C_3-C_4 intermediates, for critically discussing my ideas and data and for trusting me with the project!

I want to thank the other members of my thesis advisory committee, namely **Prof. Dr. Stan Kopriva** and **Prof. Dr. Benjamin Stich** for their constant input and support since grad school.

I owe deep thanks to all my collaborators and people I had the chance to work with over the last years. I want to pick out a few, **Dr. Ali Denton** for patiently guiding me through the quirks of bioinformatics and trusting me all the server admin work (yeah, thanks…), **Dr. Jan Buchmann** who encouraged me to follow the transposon bread crumbs and **Gabi Buchmann** without whom I'd still be in the lab trying to assemble a WGBS library! Thanks to **Sam Flachbart** for running my samples on the BioAnalyzer and to **Dr. Tobias Jores** for great discussions about enhancer elements and trusting me his unpublished STARR-seq data.

Thanks to all the people I met through CEPLAS, especially the 2019 grad school and a few special people from the CEPLAS office, **Dr. Céline Hönl**, **Dr. Nadine Rademacher**, **Dr. Petra Bracht** and **Brigitte Haumann** for the financial support, letting my hijack the CEPLAS Instagram account and being brave enough to hand me a microphone on public events.

Thanks to my students, especially **Noah**, **Krzysztof**, **Amelie**, **Rylee**, **Jacqueline**, **Celina**, **Carina** and **Patrick** for their help and trust.

I am incredibly grateful that some colleagues became real friends. I owe **Dr. Vanessa Reichel-Deland** more than I'll admit to her, thanks for keeping me sane, for shared laughs, tears, Schneehäschen, rice dishes in boats, beer, fun and not-so-fun moments. And nuclei!

Thanks to the WeberLab cool kids, **Vanessa**, **Karo**, **Jose**, several **Hanna(h)s**, **Lucas**, **Luca**, **Laia, Steffi** and everyone that I forgot to mention. **Jan, Moritz** and **Dr. Vanessa** were so kind to proofread this thesis. Thanks!

Finally, none of this would be ever possible with my family that gives me constant and unconditional support. Thanks to **Jana** for emotional support, (really) always being there and being the best person that I have ever met. My deepest thanks go to **Anni**, **Günther**, **Gudrun**, **Engelbert**, **Julia**, **Louisa** and **Johannes**, who have made and continue to make me what I am. This thesis is dedicated to you.

4

Preface

Natural variation is a gift in any endeavor to optimize plant traits. In breeding, natural variation has been used for centuries to increase the efficiency and resilience of crops. In fundamental science, the occurrence of variation in certain traits can help to unravel their genetic foundations. A promising route to enhance plant photosynthesis is to utilize natural variation in carbon assimilation pathways. A major bottleneck in photosynthetic carbon assimilation is caused by the affinity of the enzyme ribulose-1,5-bisphoshate carboxylase/oxygenase (Rubisco) to molecular oxygen, leading to the formation of a toxic metabolite that needs to be converted during photorespiration, an energy-expensive pathway. Mechanisms that decrease the need for photorespiration by minimizing the Rubisco oxygenation reactions occur naturally in carbon concentrating mechanisms such as CAM and C_4 plants. C_4 photosynthesis is a highly complex trait and efforts to implement C_4 photosynthesis into major crops have so far only led to limited success. In this work, we make use of plants exhibiting C_3-C_4 intermediate photosynthesis to learn the genetic principles behind the evolution of complex carbon concentrating traits. Specifically, we focus on the Brassicaceae, a family containing closely related crop and model species.

This thesis begins with an introduction, where I give a brief overview over the research motivation and highlight the role of *cis*-regulatory elements in the evolution of complex traits. The introduction is followed by three manuscripts that form the main body of the thesis. These span a holistic view on the single-cell transcriptome of a C_3-C_4 intermediate plant to the functional analysis of causal loci. My contribution is indicated on the first page of each manuscript.

In Manuscript I, we analyze the differential gene expression between the C_3 model plant *Arabidopsis thaliana* and the C_3-C_4 intermediate plant *Moricandia arvensis* at single-cell resolution. We find a significant recruitment of gene expression to bundle-sheath cells and report the shift of multiple components involved in photorespiratory glycine decarboxylation to this cell type. Furthermore, we shed light on the gene regulatory framework of cell-specific expression using the integration of experimental data and predictions from machine learning approaches.

In Manuscript II, published in *Plant Biology*, we use a panel of 15 Brassicaceae species, consisting of C_3 species and a polyphyletic clade of C_3-C_4 intermediate species. Within this panel, we perform a genome wide association study to correlate the presence of upstream transposable elements to C_3-C_4 intermediate traits. Among others, we find a highly correlating transposable element insertion in the upstream region of the *GLDP1* gene that is known to be differentially expressed between C_3 and C_3-C_4 intermediate species.

As a follow-up, Manuscript III focuses on the functional validation of a proposed mechanism by which the transposable element in the *GLDP1* upstream region mediates bundle-sheath cell specific expression. Using sequence comparisons, DNA methylation sequencing and promoter truncation we try to identify functional elements in the *GLDP1* promoter and hypothesize that natural structural and epigenetic variation led to a shift in the *GLDP1* expression domain, which is a key event in the evolution of C_3-C_4 intermediate traits in the Brassicaceae. Finally, the thesis closes with an outlook on new research prospects to study the genetic framework of C_3-C_4 intermediacy in closer detail.

5

Table of Contents

6

Introduction

Photosynthesis in the light of global challenges

In the 21st century, humanity faces pressing global concerns such as anthropogenic climate change and a growing population projected to exceed 10 billion individuals by the 2080s (United Nations World Population Prospects). Global food demand is expected to increase by up to 56% between 2010 and 2050 (van Dijk *et al.*, 2021), which collides with the availability of cultivable land. Concurrently, disparities in land use efficiency, notably pronounced in economically disadvantaged regions, compound this challenge significantly (Duro *et al.*, 2020). Achieving sustainable solutions under these circumstances is highly complex. As many economically important crops have already reached their theoretical yield maxima, a so-far underestimated but much-discussed attempt to increase crop yield is the optimization of photosynthesis (reviewed in Ort *et al.* (2015)). Improving photosynthesis has the potential to enhance plant resilience to fluctuating environmental influences, potentially assuring a path towards sustainable food production. The term “improving” may have different connotations from varying fields of research. In the context discussed here, it can be defined as enhancing the crop plants’ capabilities to utilize captured light energy and CO_2 more efficiently (Theeuwen *et al.*, 2022).

Terrestrial life depends fundamentally upon oxygenic photosynthesis, serving as the entry point for atmospheric CO_2 into biomass. Oxygenic photosynthesis comprises two sets of reactions. The first set, known as the light-dependent reactions, occurs within the chloroplast thylakoid membranes of plants. Here, light energy harvested by photosystem II facilitates the separation of electrons from water molecules, resulting in the release of O_2 and H^+. These molecules are then used to produce energy-rich nicotinamide adenine dinucleotide phosphate (NADPH) and adenosine triphosphate (ATP). The second set of reactions, referred to as the Calvin-Benson-Bassham (CBB) Cycle, is light-independent. These reactions utilize the NADPH and ATP generated during the light-dependent reactions to fix CO_2 into energy-rich compounds, notably glucose and other essential metabolites. Through this process, terrestrial plants assimilate over 100 billion tons of CO_2 annually, playing a crucial role in the global carbon cycle (Baslam *et al.*, 2020).

The initial fixation of inorganic CO_2 into organic biomolecules is catalyzed by the enzyme Ribulose-1,5-bisphosphate carboxylase/oxygenase (Rubisco). In the reaction, CO_2 is bound to the five-carbon compound ribulose-1,5-bisphosphate, forming two molecules of 3-phosphoglycerate (3-PGA). Within the CBB, five out of six molecules of 3-PGA undergo recycling to ribulose-1,5-bisphosphate, while one molecule of 3-PGA can be utilized to synthesize starch, sucrose or cellulose or can be directed towards glycolysis. Rubisco also catalyzes a reaction between ribulose-1,5-bisphosphate and O_2 instead of CO_2, which leads to the formation of 2-phosphoglycolate (2-PG), a potent inhibitor of core enzymes in plant metabolism. To counteract this, 2-PG is constantly and swiftly detoxified by the plant in a process called photorespiration, expending significant energy, nitrogen and carbon resources to recycle 2-PG into 3-PGA (reviewed in Bauwe *et al.* (2010)).

Utilizing natural variation of photosynthesis

Although plants occupy a wide spectrum of dynamic environments and ecological niches, their core photosynthetic machinery, encompassing around 100 genes, remains remarkably conserved (Theeuwen *et al.*, 2022). Variation in photosynthesis manifests mainly in the form of carbon concentrating mechanisms (CCMs) such as C_4 and crassulacean acid metabolism (CAM) photosynthesis. CCMs increase the partial pressure of CO_2 around the Rubisco active center and, in doing so, reduce the ratio of oxygenation to carboxylation reactions. By this, energy-expensive photorespiration is reduced, allowing plants to close their stomata under hot and dry climate conditions, making both C_4 and CAM photosynthesis highly water-use efficient. C_4 and CAM photosynthesis rely on a separation between the initial CO_2 fixation and the Rubisco reaction. In CAM photosynthesis, CO_2 uptake through stomata is performed at night and the carbon is stored as malate. Malate is degraded during the day and the released CO_2 is used in the Rubisco reaction. Contrasting to this temporal separation, C_4 photosynthesis works using a spatial separation between the initial CO_2 fixation in the mesophyll cells (MC) and the Rubisco reaction in bundle-sheath cells (BSC). In C_4 photosynthesis, CO_2 is initially fixed in the MC by phosphoenolpyruvate carboxylase and an aspartate aminotransferase or a malate dehydrogenase, depending on the C_4 photosynthesis subtype. The resulting aspartate or malate then diffuses to the BSC, where it is decarboxylated and the resulting CO_2 is used in an oxygen-depleted environment by the BSC Rubisco.

Exploiting the high water- and nitrogen-use efficiency of C_4 crops such as sugarcane, sorghum and maize has been proposed for a transformative “second green revolution” (von Caemmerer *et al.*, 2012). However, the anatomical and biochemical traits that shape C_4 photosynthesis are highly complex and no functioning C_4 cycle was so far introduced in C_3 crops (reviewed in Schuler *et al.* (2016)). Engineering a C_4 cycle into important crop plants such as rice would require a functionalization of BSC with a distinct Kranz anatomy, a reduction of vein spacing, over 100-fold increase in expression strength of core photosynthesis genes and differential partitioning of gene expression between MC and BSC tissue (Furbank *et al.*, 2023). To date, some progress has been made in BSC activation, meaning increasing photosynthetic activity in the BSC by increasing organelle number and size (Sage *et al.*, 2014). Moreover, by overexpressing GOLDEN2 and GOLDEN2-LIKE transcription factors, some traits of proto-Kranz BSC anatomy were recreated in rice (Wang *et al.*, 2017; Ermakova *et al.*, 2020).

Recent advances in synthetic biology methodologies have shown promise by increasing the overall expression strength of individual genes. On the other side, only few genetic elements responsible for spatially differential expression of genes between MC and BSC were so far detected and their role in gene-regulatory networks remains elusive (Wiludda *et al.*, 2012; Gowik *et al.*, 2017; Dickinson *et al.*, 2023). Finding the DNA motifs that play roles in the cell-specific expression of photorespiratory and C_4 genes is of pivotal interest since the differential partitioning of photorespiratory genes seems to be one of the early steps in the evolution of C_4 photosynthesis (Sage *et al.*, 2012; Heckmann *et al.*, 2013; Sage *et al.*, 2014).

It was hypothesized that the evolution of C_4 photosynthesis followed a step-wise (Sage *et al.*, 2012; Sage *et al.*, 2014) or smooth (Heckmann *et al.*, 2013) landscape of sequential biochemical and anatomical adjustments. More than 50 species in 11 families from eudicot and monocot lineages were discovered that represent stable intermediate species on the evolutionary track towards C_4 photosynthesis (Sage *et al.*, 2018; Lundgren, 2020). These so-called C_3-C_4 intermediate or C_2 species prevail on all continents except Antarctica in diverse environments and climate conditions with a tendency towards growth in warm habitats, a trend that may have facilitated the evolution of a complete C_4 syndrome. This is reinforced by the observation that clades containing C_3-C_4 intermediates but lacking C_4 species tend to expand into rather wet, less hot environments with richer soils (Lundgren *et al.*, 2017). The repeated convergent evolution of a C_3-C_4 intermediate photosynthesis trait provides an opportunity to dissect the fundamental genomics of the trait (Edwards, 2019).

C_3-C_4 intermediate photosynthesis is a collection of physiological, anatomical and genetic traits that evolved independently but show remarkable convergence. Despite this convergence, not all traits are uniformly shared among C_3-C_4 intermediate species, leading to a diverse range of "C_3-ness" within this category. One of the characteristic traits of C_3-C_4 intermediate species is a lowered carbon compensation point (CCP), which can be used to distinguish C_3-C_4 plants from C_3 species and quantify the degree of “C_3-ness” of the species’ physiology. Other physiological traits of C_3-C_4 intermediate species include a higher net photosynthetic assimilation e.g. in intermediate *Diplotaxis* and *Heliotropium* species (Ueno *et al.*, 2006; Vogan *et al.*, 2007) and a higher water-use efficiency in intermediate *Cleome* species (Voznesenskaya *et al.*, 2007). In some species, inducible C_4 photosynthesis was reported from the C_3-C_4 intermediate state, triggered by low CO_2 in *Steinichsma* (Studer, unpublished) or low nitrogen availability in *Chenopodium* (Oono *et al.*, 2022).

The physiological traits observed in C_3-C_4 intermediate species are based on a distinct leaf anatomy and biochemistry. One example is the characteristic Kranz-like leaf anatomy, in which sometimes enlarged BSC show an accumulation of chloroplasts along the cell wall towards the leaf vein. This phenomenon is observed in several, though not all C_3-C_4 intermediate species, frequently accompanied by an increased vein density (Sage *et al.*, 2014). In C_3-C_4 intermediate *Moricandia* species, such an increased vein density could not be detected, but individual veins appeared thicker, putatively due to the large number of chloroplasts arranged centripetally around the veins (Schlüter *et al.*, 2016; Schlüter *et al.*, 2017).

Whereas anatomical traits exhibit high variation between C_3-C_4 intermediate species, their biochemistry shows convergence. A common feature between multiple C_3-C_4 intermediate species is the installation of a photorespiratory glycine shuttle, which shifts a substantial fraction of photorespiration to the BSC. By restriction of the photorespiratory glycine decarboxylation to the BSC, the CO_2 released during this step is trapped inside the BSC, where it can be refixed by Rubisco (Monson *et al.*, 1984). The relocation of photorespiration by restricting glycine decarboxylation to the BSC was predicted to be the first (Heckmann *et al.*, 2013) or one of the midway (Sage *et al.*, 2012) steps in C_4 evolution. The localization of proteins from the glycine decarboxylase complex (GDC) and the serine hydroxymethyltransferase (SHMT) specifically within the BSC mitochondria in C_3-C_4 intermediate plants was discovered early in the history of C_3-C_4 photosynthesis research and led to the first models of a photorespiratory glycine shuttle (Monson *et al.*, 1984; Rawsthorne *et al.*, 1988) (Fig. 1). It is striking that the installation of a glycine shuttle evolved convergently across multiple plant taxa by restricting GDC proteins to the BSC, especially the GDC P-protein (GDC-P) (Khoshravesh *et al.*, 2016; Schulze *et al.*, 2016).

The restriction of glycine decarboxylation to the BSC leads to the release of CO_2 but also NH_3 in the BSC. Thus, rebalancing of nitrogen is required between the MC and BSC and putative pathways for this were proposed using flux modeling. The models suggest a glutamate–2-oxoglutarate, alanine–pyruvate, and aspartate–malate shuttle, parts of which resemble a primordial C_4 metabolism (Mallmann *et al.*, 2014). This is perceived as an evolutionary driving force towards C_4 photosynthesis and, supporting this, over 90% of plant tribes containing C_3-C_4 intermediate species also contain C_4 plants (Sage *et al.*, 2011; Heckmann *et al.*, 2013; Mallmann *et al.*, 2014).

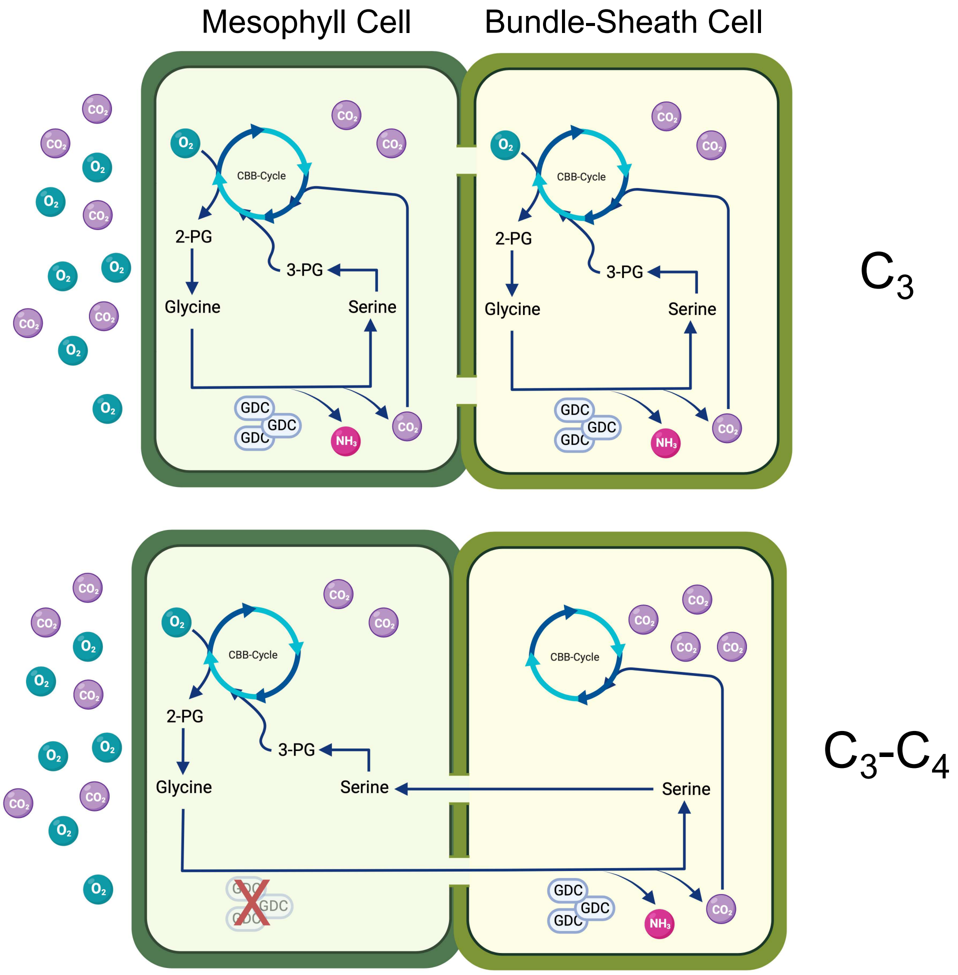


Figure 1: Simplified schematic illustration of the photorespiratory pathway in C_3 and C_3-C_4 intermediate species with a glycine shuttle. In the C_3 state, photorespiration takes place in both mesophyll cells (MC) and bundle-sheath cells (BSC) independently. In C_3-C_4 intermediate plants with a photorespiratory glycine shuttle, reactions catalyzed by the glycine decarboxylase complex (GDC) only take place in the BSC, leading to a BSC-selective release of CO_2 and NH_3. This in turn creates a high CO_2 partial pressure around Rubisco in the BSC and a net decreased number of oxygenation reactions. Abbreviations: 3-PG: 3-phosphogylcerate; 2-PG: 2-phosphoglycolate; CBB: Calvin-Benson-Bassham.

Known genetic principles of C_3-C_4 intermediate photosynthesis

Shifting the expression of a gene in space or time is one of the basic principles in the evolution of any complex trait such as C_3-C_4 intermediate photosynthesis. Throughout the evolutionary pathway towards C_3-C_4 photosynthesis, innovation primarily stems from the reconfiguration and interconnection of existing genetic modules, rather than the invention of novel genes or pathways (Hibberd *et al.*, 2010; Westhoff *et al.*, 2010; Lyu *et al.*, 2021). The establishment of the photorespiratory glycine shuttle is exemplary of this, since it only requires a spatial expression shift to prime a decisive step in C_3-C_4 evolution. Multiple studies reinforced the idea of rewiring genetic networks in C_4 evolution. These studies described the establishment of a link between C_4 gene expression and light perception networks (Singh *et al.*, 2023), the assimilation of C_4 genes into ancestral networks (Swift *et al.*, 2023) and the interplay between C_3 and C_4 gene regulatory networks at different stages of C_4 evolution (Lyu *et al.*, 2021; Lyu *et al.*, 2022). All studies reported a remarkable conservation of gene regulatory networks across a range of species and across C_3, C_3-C_4 intermediate and C_4 species. It is therefore plausible that the reconfiguration of these networks required only few genetic adjustments (Westhoff *et al.*, 2010).

The linkage between gene regulatory networks often occurs through the recruitment of *cis*-regulatory elements (CREs), which are DNA sequences harboring enhancers, insulators, or silencer elements. Various mechanisms facilitate the recruitment of CREs to a gene, for instance through the mutation of existing sequences (Behrens *et al.*, 2010) or the exaptation of transposable elements (Feschotte, 2008). Upon recruitment of a new CRE, a gene may gain or lose spatial or temporal expression characteristics. To be effective in evolution, in most cases, this neofunctionalization must not preclude the original function of the gene, which could lead to decreased fitness of the host organism. Oftentimes, neofunctionalization by co-option of new CREs is accompanied by gene duplications, a concept that was also discussed for C_4 photosynthesis traits (Monson *et al.*, 2000). Variations in gene or even genome copy number are therefore strongly connected with the genetic space that allows evolution of complex traits.

Using the Brassicaceae family to study C_3-C_4 intermediate photosynthesis

An interesting system to study the genomic foundations of C_3-C_4 intermediate photosynthesis is the Brassicaceae family. This family comprises about 4000 species including economically important crops such as arugula, canola, cabbage and radish as well as model species such as *Arabidopsis thaliana*.

Within the Brassicaceae family no C_4 plants have yet been found, but five independent evolutionary origins of C_3-C_4 intermediate photosynthesis were identified (Sage *et al.*, 2011; Guerreiro *et al.*, 2023). Whereas C_3 Brassicaceae species exhibit CCPs from 40-60 ppm, their C_3-C_4 Brassicaceae sister species exhibit notably lower CCPs, reaching around 10 ppm in *Diplotaxis tenuifolia* (arugula). Furthermore, C_3-C_4 intermediate Brassicaceae species have a centripetal arrangement of BSC organelles and, in general, higher glycine and serine levels in the leaves (Schlüter *et al.*, 2023). In contrast to other study systems such as *Flaveria* species, where C_4-like intermediate species with a weak C_4 cycle exist, Brassicaceae C_3-C_4 intermediate species exhibit a comparatively lower degree of "C_4-ness".

The possibility to study the complex suite of C_3-C_4 intermediate traits at their early stage of evolution, the repeated independent evolution of the phenotype, the availability of numerous genome sequences and the economical and scientific importance of the species make the Brassicaceae family a highly promising research system. The analysis of these nuclear and plastid genomes has provided initial insights into the phylogenetic relationships among C_3 and C_3-C_4 intermediate Brassicaceae species. It confirmed the presence of at least five independent evolutionary origins of C_3-C_4 intermediacy in this family, as well as the presence of natural hybrids between C_3 and C_3-C_4 intermediate *Diplotaxis* species (Garassino *et al.*, 2022; Guerreiro *et al.*, 2023). Throughout evolution, the Brassicaceae family underwent numerous instances of whole-genome duplications, resulting in polyploid genomes. Facilitating the neofunctionalization of duplicated genes, these polyploid genomes provide the essential genome copy number variation for the evolution of complex traits (Walden *et al.*, 2020). Interestingly, all C_3-C_4 intermediate Brassicaceae species can be found in the Brassiceae tribe (Koch *et al.*, 2012; Guerreiro *et al.*, 2023). This tribe originated around 24.2 million years ago (Arias *et al.*, 2014) and exhibits high net diversification rates. The Brassicaceae family underwent two ancient genome duplications, the At-β event shared by Brassicaceae and Cleomaceae around 170–235 million years ago and the At-α event shared by all Brassicaceae around 40 million years ago (Barker *et al.*, 2009; Walden *et al.*, 2020). Species in Brassiceae tribe, containing the C_3-C_4 intermediate species but not the model genus *Arabidopsis*, underwent an additional triplication event (Br-α) that most likely occurred through genome duplication and the addition of a third genome, leading to a hexaploidization event (Lysak *et al.*, 2009; Wang *et al.*, 2011). These numerous genome rearrangements might have been a critical prerequisite for evolution of C_3-C_4 intermediacy in this tribe.

Despite Brassicaceae species being in the center of attention in plant research, many questions are still open regarding variation in C_3-C_4 intermediate photosynthesis in this family. It is unknown why the evolutionary model of a primordial C_4 cycle as a consequence of photorespiratory nitrogen imbalance does not lead to the presence of C_4 Brassicaceae species. It is further unknown how the leaf anatomy contributes to the C_3-C_4 intermediate traits and which genetic regulators underly this. So far, the models of a Brassicaceae C_3-C_4 intermediate metabolism only rely on the differential localization of the GDC P-protein in C_3-C_4 intermediate *Moricandia*, but no further differential gene activity between MC and BSC was shown. This thesis aims to paint a more detailed picture of C_3-C_4 intermediate photosynthesis in Brassicaceae, showing the expression differences between MC and BSC as well as the underlying functional genetic and epigenetic mechanisms.

Literature

Arias, T., Beilstein, M. A., Tang, M., McKain, M. R. and Pires, J. C. (2014) 'Diversification times among Brassica (Brassicaceae) crops suggest hybrid formation after 20 million years of divergence', *American Journal of Botany,* 101(1), pp. 86-91.

Barker, M. S., Vogel, H. and Schranz, M. E. (2009) 'Paleopolyploidy in the Brassicales: Analyses of the Cleome Transcriptome Elucidate the History of Genome Duplications in Arabidopsis and Other Brassicales', *Genome Biology and Evolution,* 1, pp. 391-391.

Baslam, M., Mitsui, T., Hodges, M., Priesack, E., Herritt, M. T., Aranjuelo, I. and Sanz-Sáez, Á. (2020) 'Photosynthesis in a Changing Global Climate: Scaling Up and Scaling Down in Crops', *Frontiers in Plant Science,* 11, pp. 515969-515969.

Bauwe, H., Hagemann, M. and Fernie, A. R. (2010) Photorespiration: players, partners and origin. *Trends in Plant Science.* Elsevier Current Trends.

Behrens, S. and Vingron, M. (2010) 'Studying the evolution of promoter sequences: A waiting time problem', *Journal of Computational Biology,* 17(12), pp. 1591-1606.

Dickinson, P. J., Triesch, S., Schlüter, U., Weber, A. P. M. and Hibberd, J. M. (2023) 'A transcription factor module mediating C2 photosynthesis', *bioRxiv*, pp. 2023.09.05.556297-2023.09.05.556297.

Duro, J. A., Lauk, C., Kastner, T., Erb, K. H. and Haberl, H. (2020) 'Global inequalities in food consumption, cropland demand and land-use efficiency: A decomposition analysis', *Global Environmental Change,* 64, pp. 102124-102124.

Edwards, E. J. (2019) 'Evolutionary trajectories, accessibility and other metaphors: the case of C4 and CAM photosynthesis', *New Phytologist,* 223(4), pp. 1742-1755.

Ermakova, M., Danila, F. R., Furbank, R. T. and von Caemmerer, S. (2020) 'On the road to C4 rice: advances and perspectives', *Plant Journal,* 101(4), pp. 940-950.

Feschotte, C. (2008) 'Transposable elements and the evolution of regulatory networks', *Nature Reviews Genetics,* 9(5), pp. 397-405.

Furbank, R., Kelly, S. and von Caemmerer, S. (2023) 'Photosynthesis and food security: the evolving story of C4 rice', *Photosynthesis Research,* 158(2), pp. 121-130.

Garassino, F., Wijfjes, R. Y., Boesten, R., M Becker, F. F., Clapero, V., van den Hatert, I., Holmer, R., Eric Schranz, M., Harbinson, J., de Ridder, D., Smit, S. and M Aarts, M. G. (2022) 'The genome sequence of Hirschfeldia incana, a species with high photosynthetic light-use efficiency', *bioRxiv*, pp. 1-16.

Gowik, U., Schulze, S., Saladié, M., Rolland, V., Tanz, S. K., Westhoff, P. and Ludwig, M. (2017) 'A MEM1-like motif directs mesophyll cell-specific expression of the gene encoding the C4 carbonic anhydrase in Flaveria', *Journal of Experimental Botany,* 68(2), pp. 311-320.

Guerreiro, R., Bonthala, V. S., Schlüter, U., Hoang, N. V., Triesch, S., Schranz, M. E., Weber, A. P. M. and Stich, B. (2023) 'A genomic panel for studying C3-C4 intermediate photosynthesis in the Brassiceae tribe', *Plant Cell and Environment,* 46(11), pp. 3611-3627.

Heckmann, D., Schulze, S., Denton, A., Gowik, U., Westhoff, P., Weber, A. P. M. and Lercher, M. J. (2013) 'Predicting C4 photosynthesis evolution: Modular, individually adaptive steps on a mount fuji fitness landscape', *Cell,* 153(7), pp. 1579-1579.

Hibberd, J. M. and Covshoff, S. (2010) 'The regulation of gene expression required for C4 photosynthesis', *Annual Review of Plant Biology,* 61, pp. 181-207.

Khoshravesh, R., Stinson, C. R., Stata, M., Busch, F. A., Sage, R. F., Ludwig, M. and Sage, T. L. (2016) 'C3–C4 intermediacy in grasses: organelle enrichment and distribution, glycine decarboxylase expression, and the rise of C2 photosynthesis', *Journal of Experimental Botany,* 67(10), pp. 3065-3078.

Koch, M. A., Kiefer, M., German, D. A., Al-Shehbaz, I. A., Franzke, A., Mummenhoff, K. and Schmickl, R. (2012) 'BrassiBase: Tools and biological resources to study characters and traits in the Brassicaceae—version 1.1', *TAXON,* 61(5), pp. 1001-1009.

Lundgren, M. R. (2020) 'C2 photosynthesis: a promising route towards crop improvement?', *New Phytologist,* 228(6), pp. 1734-1740.

Lundgren, M. R. and Christin, P. A. (2017) 'Despite phylogenetic effects, C3-C4 lineages bridge the ecological gap to C4 photosynthesis', *Journal of experimental botany,* 68(2), pp. 241-254.

Lysak, M. A., Koch, M. A., Beaulieu, J. M., Meister, A. and Leitch, I. J. (2009) 'The dynamic ups and downs of genome size evolution in Brassicaceae', *Molecular Biology and Evolution,* 26(1), pp. 85-98.

Lyu, M.-J. A., Gowik, U., Kelly, S., Covshoff, S., Hibberd, J. M., Sage, R. F., Ludwig, M., Wong, G. K.-S., Westhoff, P. and Zhu, X.-G. (2021) 'The coordination of major events in C4 photosynthesis evolution in the genus Flaveria', *Scientific Reports,* 11(1), pp. 1-14.

Lyu, M. J., Tang, Q., Wang, Y., Essemine, J., Chen, F., Ni, X., Chen, G. and Zhu, X. G. (2022) 'Evolution of gene regulatory network of C4 photosynthesis in the genus Flaveria reveals the evolutionary status of C3-C4 intermediate species', *Plant Communications*, pp. 100426-100426.

Mallmann, J., Heckmann, D., Bräutigam, A., Lercher, M. J., Weber, A. P. M., Westhoff, P. and Gowik, U. (2014) 'The role of photorespiration during the evolution of C4 photosynthesis in the genus Flaveria', *eLife,* 2014(3), pp. 1-23.

Monson, R. K. and Edwards, G. E. (1984) 'C3 - C4 Intermediate Photosynthesis in Plants', *BioScience,* 34(9), pp. 563-574.

Monson, R. K. and Rawsthorne, S. (2000) 'CO2 Assimilation in C3-C4 Intermediate Plants', pp. 533-550.

Oono, J., Hatakeyama, Y., Yabiku, T. and Ueno, O. (2022) 'Effects of growth temperature and nitrogen nutrition on expression of C3–C4 intermediate traits in Chenopodium album', *Journal of Plant Research*.

Ort, D. R., Merchant, S. S., Alric, J., Barkan, A., Blankenship, R. E., Bock, R., Croce, R., Hanson, M. R., Hibberd, J. M., Long, S. P., Moore, T. A., Moroney, J., Niyogi, K. K., Parry, M. A. J., Peralta-Yahya, P. P., Prince, R. C., Redding, K. E., Spalding, M. H., Van Wijk, K. J., Vermaas, W. F. J., Von Caemmerer, S., Weber, A. P. M., Yeates, T. O., Yuan, J. S. and Zhu, X. G. (2015) 'Redesigning photosynthesis to sustainably meet global food and bioenergy demand', *Proceedings of the National Academy of Sciences of the United States of America,* 112(28), pp. 8529-8536.

Rawsthorne, S., Hylton, C. M., Smith, A. M. and Woolhouse, H. W. (1988) 'Distribution of photorespiratory enzymes between bundle-sheath and mesophyll cells in leaves of the C3-C4 intermediate species Moricandia arvensis (L.) DC', *Planta,* 176(4), pp. 527-532.

Sage, R. F., Christin, P. A. and Edwards, E. J. (2011) 'The C 4 plant lineages of planet Earth', *Journal of Experimental Botany,* 62(9), pp. 3155-3169.

Sage, R. F., Khoshravesh, R. and Sage, T. L. (2014). From proto-Kranz to C4 Kranz: Building the bridge to C 4 photosynthesis. *Journal of Experimental Botany.*

Sage, R. F., Monson, R. K., Ehleringer, J. R., Adachi, S. and Pearcy, R. W. (2018) 'Some like it hot: the physiological ecology of C4 plant evolution', *Oecologia 2018 187:4,* 187(4), pp. 941-966.

Sage, R. F., Sage, T. L. and Kocacinar, F. (2012) 'Photorespiration and the evolution of C4 photosynthesis', *Annual Review of Plant Biology,* 63, pp. 19-47.

Schlüter, U., Bouvier, J. W., Guerreiro, R., Malisic, M., Kontny, C., Westhoff, P., Stich, B. and Weber, A. P. M. (2023) 'Brassicaceae display variation in efficiency of photorespiratory carbon-recapturing mechanisms', *Journal of Experimental Botany,* 74(21), pp. 6631-6649.

Schlüter, U., Bräutigam, A., Gowik, U., Melzer, M., Christin, P. A., Kurz, S., Mettler-Altmann, T. and Weber, A. P. M. (2017) 'Photosynthesis in C3-C4 intermediate Moricandia species', *Journal of Experimental Botany,* 68(2), pp. 191-206.

Schlüter, U. and Weber, A. P. M. (2016) 'The Road to C4 Photosynthesis: Evolution of a Complex Trait via Intermediary States', *Plant and Cell Physiology,* 57(5), pp. 881-889.

Schuler, M. L., Mantegazza, O. and Weber, A. P. M. (2016) Engineering C4 photosynthesis into C3 chassis in the synthetic biology age. *Plant Journal.*

Schulze, S., Westhoff, P. and Gowik, U. (2016) 'Glycine decarboxylase in C3, C4 and C3–C4 intermediate species', *Current Opinion in Plant Biology,* 31, pp. 29-35.

Singh, P., Stevenson, S. R., Dickinson, P. J., Reyna-llorens, I., Tripathi, A., Reeves, G., Schreier, T. B. and Hibberd, J. M. (2023) 'C 4 gene induction during de-etiolation evolved through changes in cis to allow integration with ancestral C 3 gene regulatory networks', 9(13).

Swift, J., Luginbuehl, L. H., Schreier, T. B., Donald, R. M., Lee, T. A., Nery, J. R., Ecker, J. R. and Hibberd, J. M. (2023) 'Single nuclei sequencing reveals C4 photosynthesis is based on rewiring of ancestral cell identity networks', *bioRxiv*, pp. 2023.10.26.562893-2023.10.26.562893.

Theeuwen, T. P. J. M., Logie, L. L., Harbinson, J. and Aarts, M. G. M. (2022) 'Genetics as a key to improving crop photosynthesis', *Journal of Experimental Botany,* 73(10), pp. 3122-3137.

Ueno, O., Wada, Y., Wakai, M. and Bang, S. W. (2006) 'Evidence from photosynthetic characteristics for the hybrid origin of Diplotaxis muralis from a C3-C4 intermediate and a C 3 species', *Plant Biology,* 8(2), pp. 253-259.

van Dijk, M., Morley, T., Rau, M. L. and Saghai, Y. (2021) 'A meta-analysis of projected global food demand and population at risk of hunger for the period 2010–2050', *Nature Food 2021 2:7,* 2(7), pp. 494-501.

Vogan, P. J., Frohlich, M. W. and Sage, R. F. (2007) 'The functional significance of C3-C4 intermediate traits in Heliotropium L. (Boraginaceae): Gas exchange perspectives', *Plant, Cell and Environment,* 30(10), pp. 1337-1345.

von Caemmerer, S., Quick, W. P. and Furbank, R. T. (2012) 'The development of C4 rice: Current progress and future challenges', *Science,* 336(6089), pp. 1671-1672.

Voznesenskaya, E. V., Koteyeva, N. K., Chuong, S. D. X., Ivanova, A. N., Barroca, J., Craven, L. A. and Edwards, G. E. (2007) 'Physiological, anatomical and biochemical characterisation of photosynthetic types in genus Cleome (Cleomaceae)', *Functional Plant Biology,* 34(4), pp. 247-267.

Walden, N., German, D. A., Wolf, E. M., Kiefer, M., Rigault, P., Huang, X. C., Kiefer, C., Schmickl, R., Franzke, A., Neuffer, B., Mummenhoff, K. and Koch, M. A. (2020) 'Nested whole-genome duplications coincide with diversification and high morphological disparity in Brassicaceae', *Nature Communications,* 11(1).

Wang, P., Khoshravesh, R., Karki, S., Tapia, R., Balahadia, C. P., Bandyopadhyay, A., Quick, W. P., Furbank, R., Sage, T. L. and Langdale, J. A. (2017) 'Re-creation of a Key Step in the Evolutionary Switch from C3 to C4 Leaf Anatomy', *Current Biology,* 27(21), pp. 3278-3287.e6.

**Wang, X. and Wang, H. and Wang, J. and Sun, R. and Wu, J. and Liu, S. and Bai, Y. and Mun, J. H. and Bancroft, I. and Cheng, F. and Huang, S. and Li, X. and Hua, W. and Wang, J. and Wang, X. and Freeling, M. and Pires, J. C. and Paterson, A. H. and Chalhoub, B. and Wang, B. and Hayward, A. and Sharpe, A. G. and Park, B. S. and Weisshaar, B. and Liu, B. and Li, B. and Liu, B. and Tong, C. and Song, C. and Duran, C. and Peng, C. and Geng, C. and Koh, C. and Lin, C. and Edwards, D. and Mu, D. and Shen, D. and Soumpourou, E. and Li, F. and Fraser, F. and Conant, G. and Lassalle, G.

and King, G. J. and Bonnema, G. and Tang, H. and Wang, H. and Belcram, H. and Zhou, H. and Hirakawa, H. and Abe, H. and Guo, H. and Wang, H. and Jin, H. and Parkin, I. A. P. and Batley, J. and Kim, J. S. and Just, J. and Li, J. and Xu, J. and Deng, J. and Kim, J. A. and Li, J. and Yu, J. and Meng, J. and Wang, J. and Min, J. and Poulain, J. and Hatakeyama, K. and Wu, K. and Wang, L. and Fang, L. and Trick, M. and Links, M. G. and Zhao, M. and Jin, M. and Ramchiary, N. and Drou, N. and Berkman, P. J. and Cai, Q. and Huang, Q. and Li, R. and Tabata, S. and Cheng, S. and Zhang, S. and Zhang, S. and Huang, S. and Sato, S. and Sun, S. and Kwon, S. J. and Choi, S. R. and Lee, T. H. and Fan, W. and Zhao, X. and Tan, X. and Xu, X. and Wang, Y. and Qiu, Y. and Yin, Y. and Li, Y. and Du, Y. and Liao, Y. and Lim, Y. and Narusaka, Y. and Wang, Y. and Wang, Z. and Li, Z. and Wang, Z. and Xiong, Z. and Zhang, Z. (2011) 'The genome of the mesopolyploid crop species Brassica rapa', *Nature Genetics 2011 43:10,* 43(10), pp. 1035-1039.

Westhoff, P. and Gowik, U. (2010) 'Evolution of C4 photosynthesis-looking for the master switch', *Plant Physiology,* 154(2), pp. 598-601.

Wiludda, C., Schulze, S., Gowik, U., Engelmann, S., Koczor, M., Streubel, M., Bauwe, H. and Westhoff, P. (2012) 'Regulation of the photorespiratory GLDPA gene in C 4 Flaveria: An intricate interplay of transcriptional and posttranscriptional processes', *Plant Cell,* 24(1), pp. 137-151.

"Any sufficiently advanced technology
is indistinguishable from magic."
Arthur C. Clarke

7

Manuscript I

Single-nuclei sequencing of *Moricandia arvensis* reveals bundle sheath cell functionalization in C_3-C_4 intermediate Brassicaceae

Single-nuclei sequencing of *Moricandia arvensis* reveals bundle-sheath cell functionalization in C_3-C_4 intermediate Brassicaceae

Sebastian Triesch[1], Vanessa Reichel-Deland[1], Tobias Lautwein[3], Urte Schlüter[1,2], Andreas P.M. Weber[1,2]

[1] Institute for Plant Biochemistry, Heinrich Heine University Düsseldorf, Germany
[2] Cluster of Excellence on Plant Sciences (CEPLAS)
[3] Genomics and Transcriptomics Laboratory, Heinrich Heine University Düsseldorf, Germany

Author contributions

S.T. conducted nuclei isolation, all single-cell RNA-seq experiments, data analysis and wrote the manuscript with input from all authors. **V.R.D.** established the nuclei isolation workflow and performed nuclei preparation. **T.L.** performed library preparation and sequencing. **U.S.** and **A.P.M.W.** conceptualized the project and supervised the work.

Acknowledgements

The authors thank Dr. Yi-Jun Kim for providing the *A. thaliana* scRNA-seq transcriptome, Samantha Flachbart for assisting in nuclei preparation, Jana Peter for assisting in plant handling and Dr. Tobias Jores for providing STARR-seq enhancer data. Computational infrastructure and support were provided by the Centre for Information and Media Technology at Heinrich Heine University Düsseldorf.

Data availability

All data can be found in an Annotated Research Context Format (ARC) at https://git.nfdi4plants.org/hhu-plant-biochemistry/2023_single_nuclei_rnaseq_moricandia
and
https://git.nfdi4plants.org/hhu-plant-biochemistry/2023_athaliana_snrnaseq_comparison
(private repositories as of July 2024).

Abstract

Spatially confined gene expression defines cell identity and is the basis for all complex plant traits. C_3-C_4 intermediate photosynthesis is one of these complex traits and plants exhibiting this specialized carbon concentrating mechanism are generally perceived as stable intermediates on the evolutionary progression from ancestral C_3 to more efficient C_4 photosynthesis. Within the Brassicaceae, a family comprising model plants and crop species, C_3-C_4 intermediate photosynthesis evolved independently multiple times. Despite being an interesting clade to study this promising array of traits, research on the C_3-C_4 intermediate genetics in the Brassicaceae was so far limited to few case studies of differentially localized proteins between mesophyll and bundle sheath cells. We here sought to make use of recent advancements in single-cell transcriptome sequencing to fill this knowledge gap and present the *bona fide* first leaf transcriptome of a C_3-C_4 intermediate Brassicaceae on cell-level resolution. To this end, we generated a single-nuclei RNA sequencing dataset for *Moricandia arvensis*, a Brassicaceae with C_3-C_4 intermediate characteristics and compared it to a publicly available single-cell transcriptome for the C_3 *Arabidopsis thaliana* leaf tissue.

Our analysis revealed a significant expression shift of genes associated with photorespiration, redox regulation and transport to the *M. arvensis* bundle-sheath. These genes were not confined to this cell type in the C_3 plant. Using detailed motif analysis and the integration of next-generation sequencing datasets, we could show an enrichment of light-responsive TGA and NAC binding sites in the upstream regions of bundle-sheath cell confined genes.

In C_3-C_4 intermediate plants, a glycine shuttle operates due to the bundle-sheath specific activity of the glycine decarboxylase complex. In contrast to other studies, we show here that all genes encoding proteins involved in glycine decarboxylation underly bundle-sheath confined expression. Based on our single-nuclei transcriptome we also

postulate an additional glycolate/glyoxylate shuttle between mesophyll and bundle-sheath cells that efficiently scavenges photorespiratory ammonium in the bundle-sheath cell.

Introduction

Complex plant traits necessitate distinct spatial gene regulation and C_3-C_4 intermediate photosynthesis is a highly distinguished trait that involves a specialized anatomy and metabolism (Schlüter *et al.*, 2016; Lundgren, 2020; Giacomello, 2021). Having evolved multiple times repeatedly in over 50 species, these C_3-C_4 intermediate species are a valuable source for research on the early steps of C_4 evolution and the genetics underlying cell-specific gene expression (Sage *et al.*, 2018; Lundgren, 2020; Walsh *et al.*, 2023). The Brassicaceae family contains at least five independent origins of C_3-C_4 intermediate photosynthesis next to well-characterized model species like *Arabidopsis thaliana* or *Arabis alpina* and economically relevant crop species such as *Brassica oleracea* (cabbage), *Brassica napus* (rapeseed) or *Diplotaxis tenuifolia* (arugula) (Guerreiro *et al.*, 2023). The close phylogenetic proximity of model-, crop- and C_3-C_4 plants makes the Brassicaceae family an ideal system to study the onset of C_4 evolution in greater detail. In the Brassicaceae genus *Moricandia*, it was shown that C_3-C_4 intermediate species have a decreased carbon compensation point (Rawsthorne *et al.*, 1988b) and an altered leaf ultrastructure including bundle sheath cells (BSC) with a high number of centripetally arranged chloroplasts (Schlüter *et al.*, 2017). It was also shown that the P-protein of the glycine decarboxylase complex (GDC) localizes specifically to the BSC mitochondria selectively in C_3-C_4 intermediate *Moricandia*, but not in closely related C_3 *Moricandia* species (Hylton *et al.*, 1988; Rawsthorne *et al.*, 1988b; Rawsthorne *et al.*, 1988a). GDC is a crucial component of photorespiration since it catalyzes the decarboxylation reaction of photorespiratory glycine to CO_2, NH_3

and 5,10-methyltetrahydrofolate which is converted to serine by a serine hydroxymethyltransferase (SHMT). The BSC-specific activity of the GDC P-protein was shown in multiple C_3-C_4 intermediate and C_4 species and it is widely agreed to be a critical step in the evolution of C_4 photosynthesis via C_3-C_4 intermediate stages (reviewed in Schulze *et al.* (2016)). The confinement of the GDC P-protein is a noteworthy example for differential gene regulation at the basis of complex traits. It was thought that the BSC-preferential localization of the GDC in C_3-C_4 intermediate Brassicaceae species causes an accumulation of glycine in the mesophyll cells (MC), which then passively diffuses to the BSC where it is selectively decarboxylated. This glycine shuttle and the associated BSC-specific release of CO_2 from this decarboxylation leads to an elevated CO_2 partial pressure around Rubisco and the photorespiratory CO_2 can be efficiently refixed (Bauwe *et al.*, 2010). Glycine decarboxylation also releases nitrogen in form of NH_3, creating a nitrogen imbalance towards the BSC (Monson *et al.*, 2000). Several routes for the back-shuttle of photorespiratory nitrogen have been proposed using flux modelling by Mallmann *et al.* (2014). They propose three scenarios for ammonium shuttles between the MC and BSC: a glutamate/2-oxoglutarate shuttle, an alanine/pyruvate shuttle, and an aspartate/malate shuttle. Since aspartate, malate, pyruvate and alanine itself are exchanged between MC and BSC in C_4 photosynthesis, the suggested nitrogen back-shuttles in C_3-C_4 intermediate plants could resemble primordial C_4 cycles and facilitate the evolution of the additional anatomical and metabolic C_4 traits. However, no evidence for the presence of these shuttles exists in *Moricandia* besides increased metabolite levels of glutamate, malate and alanine in C_3-C_4 intermediate *Moricandia arvensis* and *Moricandia suffruticosa*.

The establishment of a glycine shuttle initiating further steps of C_4 evolution was predicted by modelling studies, suggesting a smooth path to C_4 (Heckmann *et al.*, 2013;

Mallmann *et al.*, 2014; Sage *et al.*, 2014; Schlüter *et al.*, 2016). Earlier conceptual models place anatomic preconditioning events like the increase of BSC organelles and their localization to the inner BSC before the shift of GSC activity to the BSC (Sage, 2016). Not much is known about the genetic regulation involved in the evolution of C_4 anatomy. In C_4 grasses such as maize and *Setaria*, the contribution of GOLDEN2-like and SCARECROW transcription factors (TFs) was shown to influence BSC anatomy and chloroplast biogenesis (Slewinski *et al.*, 2012; Lambret-Frotte *et al.*, 2024) but the target genes controlled by these TFs and their cell-specific regulation remain unknown. The significant influence of TFs on the complex development of C_4 anatomy indicates that multiple genes are affected by the evolution of C_4 anatomy and that this evolution is likely mediated by changes in *cis*-regulatory factors. In this context, research on C_3-C_4 intermediate plants is valuable because these species likely harbor the genetic variation underlying the evolutionary processes in different degrees of complexity.

We assumed that to understand the genetic principles of C_3-C_4 intermediacy, we need to illuminate gene expression dynamics in a spatial context, preferably the level of the fundamental organismal unit, the single cell. Methods for single-cell transcriptome studies in C_3-C_4 intermediate or C_4 species were so far limited by the physical separation of BSC from phloem tissue or even BSC from MC (Aubry *et al.*, 2014; Burgess *et al.*, 2019; Borba *et al.*, 2023). Until now, laser microdissection has been considered as the gold standard for isolating MC and BSC in C_4 plants. This method has provided valuable insights into the C_4-specific transcriptional control at the single-cell level (Hua *et al.*, 2021; Liu *et al.*, 2022; Moreno-Villena *et al.*, 2022). Isolation of nuclei tagged in specific cell types (INTACT) allows cell-type specific transcriptome sequencing, but requires the generation of transgenic material, which is not feasible in all plant species (Deal *et al.*, 2020). Recent progress in droplet-based single-cell or single-nuclei RNA sequencing (sc/snRNA-seq) method development opened the door to the discovery of cell-specific transcriptome patterns in unprecedented resolution (reviewed in Giacomello (2021)). However, these methods have so far only been applied to few C_4 plants (Swift *et al.*, 2023) and there is *bona fide* no scRNA-seq study for C_3-C_4 intermediate Brassicaceae to date.

To address this knowledge gap, we conducted snRNA-seq on leaf tissue from the C_3-C_4 intermediate plant *M. arvensis*. In doing so, we established a nuclei isolation protocol for *M. arvensis* and utilized droplet-based snRNA-seq employing the *10X* workflow. The dataset enabled us to distinguish specific cell types and their corresponding transcriptional profiles. To ascertain that the patterns we found were associated with C_3-C_4 intermediate traits, we compared the dataset with a publicly available *A. thaliana* leaf scRNA-seq dataset (Kim *et al.*, 2021).

We explored the datasets with regard to three main research questions: (1) What role does the BSC play in a leaf with C_3-C_4 intermediate photosynthesis compared to a C_3 BSC? (2) How is the glycine shuttle integrated into the BSC-specific sulfur and nitrogen metabolism in C_3-C_4 intermediate *M. arvensis*? (3) Can we infer *cis*-regulatory patterns underlying BSC-specific gene expression?

In doing so, we examined the expression of genes involved in photorespiration and the proposed routes for transporting photorespiratory nitrogen. In addition to the glycine shuttle, our findings suggest a simple glyoxylate/glycolate shuttle to the BSC, potentially reducing the requirement for nitrogen back transport to the MC. This hypothesis could shed light on the evolutionary stability of C_3-C_4 intermediates within the Brassicaceae. The comparison of *M. arvensis* and *A. thaliana* single-cell transcriptomes also showed a remarkable recruitment of gene expression to the BSC from *A. thaliana* to *M. arvensis*. We found this BSC functionalization largely driven by the shift of genes associated with photorespiration, redox balancing and transport to the *M. arvensis* BSC. It is generally believed that the evolution of C_4 photosynthesis via C_3-C_4 intermediate stages involves the reconnection of existing genes into modified gene regulatory networks, likely *via* the exaptation of *cis*-elements (Hibberd *et al.*, 2010). To this end, we integrated multiple next generation sequencing (NGS) datasets and analyzed *M. arvensis* upstream DNA sequences for the presence of motifs for co-expressed TFs. Here, although limited by sequencing depth, we found the BSC-specific expression and light-dependent behavior of TGA and NAC TFs, which potentially mediate BSC specificity in C_3-C_4 intermediate *M. arvensis*.

Material and Methods

Plant growth

Moricandia arvensis MOR1 plants were sterilized using chlorine gas sterilization for 1 h and germinated on ½ strength Murashige and Skoog medium with 0.4 % (w/v) agar for 7 days at room temperature and a 12 h light-dark cycle. Germinated seedlings were transferred to soil and grown at a 12 h light-dark cycle and 25 °C (light) and 22 °C (dark).

Nuclei extraction

2 g leaf material of the combined 5th and 6th leaf (counting from the first cotyledon) were cut using scissors and placed on a petri dish containing 1 mL LB01 buffer (15 mM Tris-HCl pH = 7.5, 2 mM EDTA, 80 mM KCl, 20 mM NaCl, 15 mM 2-mercaptoethanol, 0.2 % (v/v) Triton X-100, 0.5 mM Spermine). Leaves were chopped on ice for 2 min with a razor blade. After addition of 1 mL LB01 buffer, leaves were chopped for another 3 min. 3 mL LB01 were added to the chopped leaves in the petri dish and the mixture was incubated on ice for 15 min with gentle agitation every 3 min. The mixture was filtered through a 100 μm filter and subsequently through a 20 μm filter. All filters were pre-wetted with 1 mL LB01. The filtrate was overlayed on density gradient

centrifugation buffer (DGCB; 1.7 M sucrose, 10 mM Tris-HCl pH = 8.0, 2 mM $MgCl_2$, 5 mM 2-mercaptoethanol, 1 mM EDTA, 0.15 % (v/v) Triton X-100) and centrifuged at 4 °C and 1500 g for 30 min. The supernatant was discarded and the pellet was resuspended with 400 μl 10X nuclei resuspension buffer. A 30 μl aliquot of the nuclei suspension was stained with 1.5 μl DAPI to check nuclei integrity under a fluorescence microscope. All buffers were supplemented with 1 U/μl RNase inhibitor (Roche).

Single Cell Library Generation

A total of 5000 cells were used as input for the single-cell droplet library generation on the 10X Chromium Controller system utilizing the Chromium Single Cell 3' NextGEM Reagent Kit v3.1 according to manufacturer's instructions. Sequencing was carried out on a NextSeq 2000 system (Illumina Inc. San Diego, USA).

Genome annotation and orthology analysis

The *A. thaliana* TAIR10 genome and annotation was obtained from www.arabidopsis.org (Lamesch *et al.*, 2012; Berardini *et al.*, 2015). The *M. arvensis* MOR1 genome was obtained from Guerreiro *et al.* (2023), the *de novo* annotation using *Helixer* (Stiehler *et al.*, 2020; Holst *et al.*, 2023*)* was obtained from Triesch *et al.* (2024), Chapter 8). *gffread 0.12.7* was used to convert gff3 to gtf files to map snRNA-seq reads. Homologs between *A. thaliana* and *M. arvensis* were identified using *MMseqs2* (Steinegger *et al.*, 2017). If multiple *M. arvensis* homologs for one *A. thaliana* gene were found, all were retained.

Processing of 10X Genomics single cell data

Raw reads from the *M. arvensis* snRNA-seq experiment were processed using *Cell Ranger 6.1.2* (10X Genomics, CA, USA) using the ARC-v1 chemistry option. The count matrix for the *A. thaliana* scRNA-seq experiment was kindly provided by Dr. Ji-Yun Kim (HHU Düsseldorf, Kim *et al.* (2021)). RNA velocity was determined using *Velocyto 0.17.17* (La Manno *et al.*, 2018) and *scvelo 0.3.1*. Further processing of the count matrix from *Cell Ranger* was performed using *scanpy 1.9.3*. Genes with annotated chloroplast and mitochondrial origin were removed. Genes occurring in less than 10 cells and cells with less than 200 genes were removed. Cells were clustered using a graph-based approach using 50 principal components and a resolution of 0.25 for *M. arvensis* and 0.5 for *A. thaliana*.

Cluster annotation and marker gene inference

Single-cell markers genes for *A. thaliana* were obtained from PlantscRNAdb (Chen *et al.*, 2021) and the *M. arvensis* homologs for these genes were used for the respective datasets. In both datasets, within each cluster, marker genes were defined as genes with a $p < 1\text{e-}20$ and a positive log2-fold change. Each marker gene was tagged with one or more cell types from the PlantscRNAdb marker gene annotation. Fisher's exact test was employed to find cell type annotation enrichments within the marker genes for the individual clusters. The enriched cell type annotation with the lowest p-value was naïvely selected for annotation of the respective cluster.

Cross-species cell-type comparison

Marker genes for each cluster/cell-type were defined as genes with $p < 1\text{e-}20$ and a positive log2fold-change. Differentially expressed genes were determined from expression matrices using Mann-Whitney-U-Tests with a significance threshold of $p < 1\text{e-}20$. For the between-species analysis, gene expression as read counts per transcript was normalized to the highest expressed marker gene of the respective cluster. *MapMan* bin classifications for enrichment analyses were taken from Triesch *et al.* (2024) (Chapter 8). *MapMan* bin enrichment was performed using Fisher's exact test, calculating the enrichment of target gene sets within the complete set of genes in the TAIR10 or *M. arvensis* annotation, respectively. Sankey plots were created using www.sankeymatic.com.

Cis-element detection and cross-species analysis

Transcription factor binding sites for all upstream sequences were predicted using *FIMO* (Grant *et al.*, 2011) from the *MEME suite 5.5.5* (Bailey *et al*., 2009). To this end, the 3000 bp upstream sequences for all annotated *M. arvensis* and *A. thaliana* genes were collected using a custom *python* script. Motif clustering was performed on transcription factor motifs using the JASPAR database (Castro-Mondragon *et al.*, 2022). To account for sequence composition a background model was generated using the *fasta-get-markov* tool from the *MEME suite 5.5.5*. The target upstream sequence, the background model and the database were used as input for *FIMO*.

Multiome data integration

To integrate data from multiple NGS experiments and predictions from machine learning software, the *M. arvensis* genome from Guerreiro *et al.* (2023) and the corresponding gene model annotation from Triesch *et al.* (2024) were used. Whole-genome bisulfite data for 5-methylcytosine DNA modifications were used from Triesch *et al*. (unpublished; Chapter 9) were mapped and quantified against the reference genome using *Bismark 0.23.1* (Krueger *et al.*, 2011). Chromatin immunoprecipitation DNA sequencing (ChIP-seq) and Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq) coverage predictions were generated using *predmoter* (Kindel *et al*., 2024). Self-transcribing active regulatory region sequencing (STARR-seq) reads were kindly provided by Dr. Tobias Jores (HHU Düsseldorf, personal communication). Core promoter STARR-seq elements were already assigned to the respective gene and were assigned to the homolog *M. arvensis* promoter using the best *MMseqs2* hit. STARR-seq enhancer elements were mapped against the reference *M. arvensis* upstream regions using *BLAST 2.6.0*. To this end, the up to 3000 bp upstream sequences of every gene annotated from *Helixer* was dumped using a custom *python* script and used as the target sequence database for *BLAST*. To visualize single-cell and multiome results, a *shiny* webserver was deployed and is accessible under https://134.99.200.66:8000 from the internal network of Heinrich Heine University Düsseldorf.

Results

Single-nuclei sequencing of *Moricandia arvensis*

To gain an understanding of the differential gene expression patterns in a leaf of a C_3-C_4 intermediate Brassicaceae species, the fifth and sixth leaves of young *M. arvensis* plants (counting from the first cotyledon) were used for nuclei extraction and snRNA-seq. Across three biological replicates, a total of 11,013 nuclei were used for snRNA-seq using the 10X single-nuclei workflow with a median of 1,070 identified genes per nucleus. The output was visualized using uniform manifold approximation projection (UMAP) and subjected to unsupervised clustering using *scanpy*, leading to ten distinct cell clusters (Fig. 1A). The expression levels of the closest *M. arvensis* homologs for single-cell marker genes from *A. thaliana* (Chen *et al.*, 2021) were used for annotation of cell types (Fig. 1B).

In the resulting dataset six mesophyll cell (MC) clusters, one bundle-sheath cell (BSC) cluster, one phloem and two epidermis cell clusters were identified. The MC clusters were identified by, among others, the strong expression of genes encoding CARBONIC ANHYDRASE homologs, subunits of photosystem 1 and RUBISCO ACTIVASE. The emergence of multiple MC subclusters could be attributed to the utilization of different biological replicates. Typically, the three biological replicates clustered together across all clusters, except for the mesophyll clusters, where the replicates exhibited more distinct clustering patterns. (Supp Figure 1). However, the six mesophyll clusters showed the same expression patterns for most genes (Supp. Data on GitLab).

Phloem clusters were identified by the expression of marker genes such as *TRX3*, *GSR2* and *ROG1*. The clusters were annotated as phloem parenchyma and phloem companion cells. Two distinct epidermis clusters were identified by expression of epidermal marker genes like *ECERIFERUM* and *KCS6* (Fig. 1B). RNA velocity analysis indicated that the two clusters represent old and young epidermal cells rather than upper and lower epidermal layers (Fig. 1E).

1,687 cells formed a cluster that could be annotated as BSC tissue. Marker genes for the BSC cluster were the sub-proteins of the glycine decarboxylase complex (GDC) as well as the *SULTR* sulfate transporter homologs. Sub-clustering of the BSC cluster using a higher clustering resolution led to two distinct BSC cell types, corresponding to anatomical findings (Schlüter *et al.*, 2017). In the latent time analysis by RNA velocity modeling, BSC from the two subclusters formed a linear gradient, suggesting that the two clusters did not originate from different developmental stages but rather from different BS cell types (Fig. 1F).

Aiming to determine the underlying genes for traits associated with C_3-C_4 intermediate photosynthesis, we calculated differential expression of genes between MC and BSC clusters using Mann-Whitney-U-Tests with a significance threshold of p < 1e-20. We found 219 genes to be differentially expressed between MC and BSC, whereas 159 genes were upregulated in BSC and 60 genes were upregulated in MC. Enrichment analysis of the 219 differentially regulated genes revealed that within these genes the *MapMan* bins “Photosynthesis” and “Solute transport” were strongly enriched. The enrichment of genes within the “Solute transport” bin was mostly due to genes encoding the A BOUT DE SOUFFLE (BOU) mitochondrial transporter, the SULTR2;2 and SULTR3;3 sulfate transporters, proton pump ATPases and transporters from the plasma membrane intrinsic protein (PIP) transporter family. The enrichment in the bin

"Photosynthesis.photorespiration" was mostly due to various components of the GDC being partitioned to the BSC. This involved the genes encoding the GDC P-, H-, and T-protein (GLDP1, GDC-H1, GLDT) as well as the gene encoding the photorespiratory serine hydroxymethyltransferase (SHM1) (Fig. 1D). Using RNA velocity analysis, a pseudo-age was modelled for each cell. Within the BSC cluster the expression of *GLDP1* was observable in cells with a higher pseudo-age, whereas expression of the mesophyll marker gene *CARBONIC ANHYDRASE 2* (*CA2*) was observed in younger BSC (Fig. 1F).

Comparative single-cell transcriptomics with *Arabidopsis thaliana*

To get insights into cell-specific expression patterns that correlate and potentially underly traits associated with C_3-C_4 intermediate photosynthesis, we compared our *M. arvensis* snRNA-seq dataset to a publicly available single-cell RNA-seq (scRNA-seq) dataset from the C_3 Brassicaceae species *Arabidopsis thaliana*. This dataset contained expression data for 5,230 vasculature-enriched leaf protoplasts from two biological replicates (Kim *et al.*, 2021). With a median of 3,342, the number of genes per cell was approximately three times higher in the *A. thaliana* protoplast dataset than in the *M. arvensis* nuclei population. To compare the datasets of both species, the *M. arvensis* homologs of *A. thaliana* genes were assigned using *MMseqs2* (Steinegger *et al.*, 2017). This approach clustered genes based on protein sequences and combined ortholog and paralog sequences. For 20,512 *A. thaliana* genes, one or more homologs could be found in *M. arvensis*. A median of two *M. arvensis* gene copies was found for each *A. thaliana* gene.

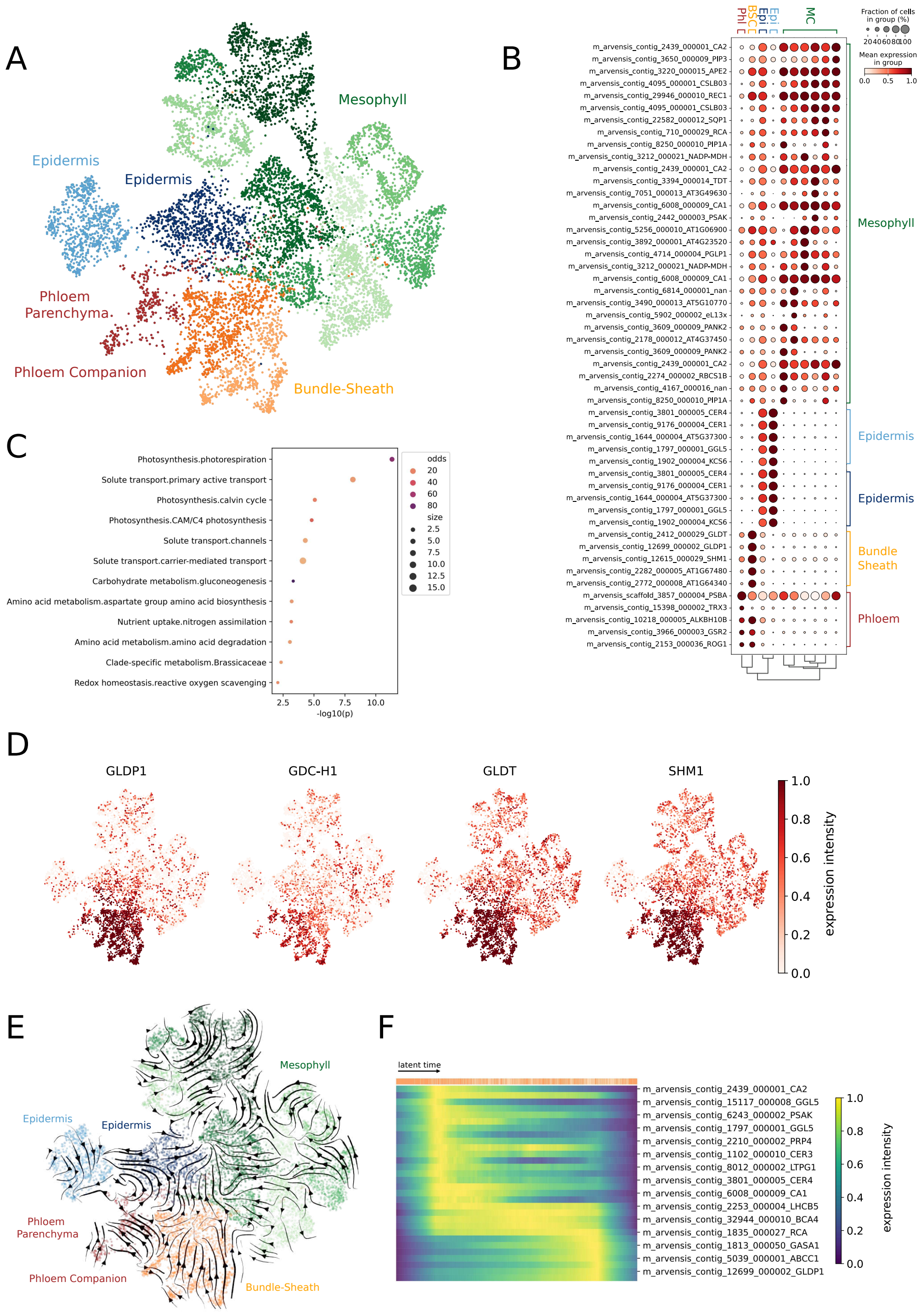


Figure 1: A: UMAP of transcript profiles for *Moricandia arvensis* 11,013 leaf cells. Cells were grouped into ten clusters using scanpy. Cell identities were assigned using marker genes for different cell types (**B**). **C: Enrichment of MapMan bins for differentially expressed genes** between *M. arvensis* bundle-sheath and mesophyll clusters. **D: UMAP with heatmap overlays indicating expression strength** of selected genes. **E: UMAP with RNA velocity arrows** modelled using velocyto and scvelo. **F: Heatmap of expression intensity** for selected genes across a modelled latent time gradient for 1,687 bundle-sheath cells.

Using the same workflow that was applied to the *M. arvensis* dataset, we visualized the *A. thaliana* scRNA-seq data using UMAP and clustered it into seven distinct clusters. However, when attempting to combine the *A. thaliana* and *M. arvensis* datasets into a single UMAP projection, the combined cells clustered solely based on their respective species, rather than according to cell types. To this end, both datasets were kept and clustered separately and comparisons were made on a per-cluster basis. The clusters were annotated using the same set of marker genes that was used for the analysis of *M. arvensis*. In contrast to the *M. arvensis* dataset, we could identify senescent, initial and guard cells in the *A. thaliana* single-cell population (Fig. 2A). Furthermore, the mean sequencing depth reached approximately 96,000 reads per cell in the *A. thaliana* dataset, whereas it was limited to around 27,000 reads per cell in our *M. arvensis* dataset.

Comparing the BSC populations between the species revealed that a proportion of *M. arvensis* BSC marker genes was not BSC-specific in *A. thaliana* (Fig. 2C). 71 *M. arvensis* BSC marker genes were also marker genes in the *A. thaliana* BSC and 101 of the *M. arvensis* BSC marker genes were not BSC-specific in *A. thaliana* (Supp. Fig. S3). The 71 shared BSC marker genes were, among others, enriched in the *MapMan* bins “Solute transport” and “Amino acid metabolism”. We referred to the 101 genes showing cell-specific expression only in one of the species as “differentially partitioned”. These genes gained BSC-specific expression in *M. arvensis* during the evolutionary progression since the last common ancestor between the two plant species. The differentially partitioned *M. arvensis* BSC marker genes were enriched in the *MapMan* bins “Photosynthesis.photorespiration”, “Solute transport.Primary active transport” and “Nutrient uptake.nitrogen assimilation”. The enrichment of “Solute transport.Primary active transport” was largely due to the BSC-specific expression of ATP-binding cassette (ABC) transporters in *M. arvensis*. The enrichment of

“Photosynthesis.photorespiration” in the set of differentially partitioned genes was due to homologs of the genes encoding subproteins of the GDC such as GLDP1, GLDT and GDC-H1 (Fig. 2B). The BSC-preferential expression also diverged between *M. arvensis* homologs of single *A. thaliana* genes. For the *GLDT* gene, for example, two copies were annotated in *M. arvensis*, one of which showed a stronger BSC-specific expression and one showed BSC-specificity with a stronger MC background. (Supp Figure S2). A similar pattern was found for the four *M. arvensis* paralogs of the single *A. thaliana* *NADP-MDH* gene, encoding the plastidial NADP-dependent malate dehydrogenase. Here, one paralog showed highly BSC-specific expression, whereas another paralog showed higher MC expression.

***Cis*-regulatory underpinnings of differential partitioning**

We next investigated the *cis*-regulatory landscape of differentially partitioned genes in the *M. arvensis* and *A. thaliana* datasets. To this end, the 3000 bp upstream regions for each marker gene were analyzed using *FIMO* (Grant *et al.*, 2011), which predicted TF binding sites. The list of putatively binding TFs based on upstream sequence elements was narrowed down based on the TF co-expression data from our single-cell RNAseq datasets.

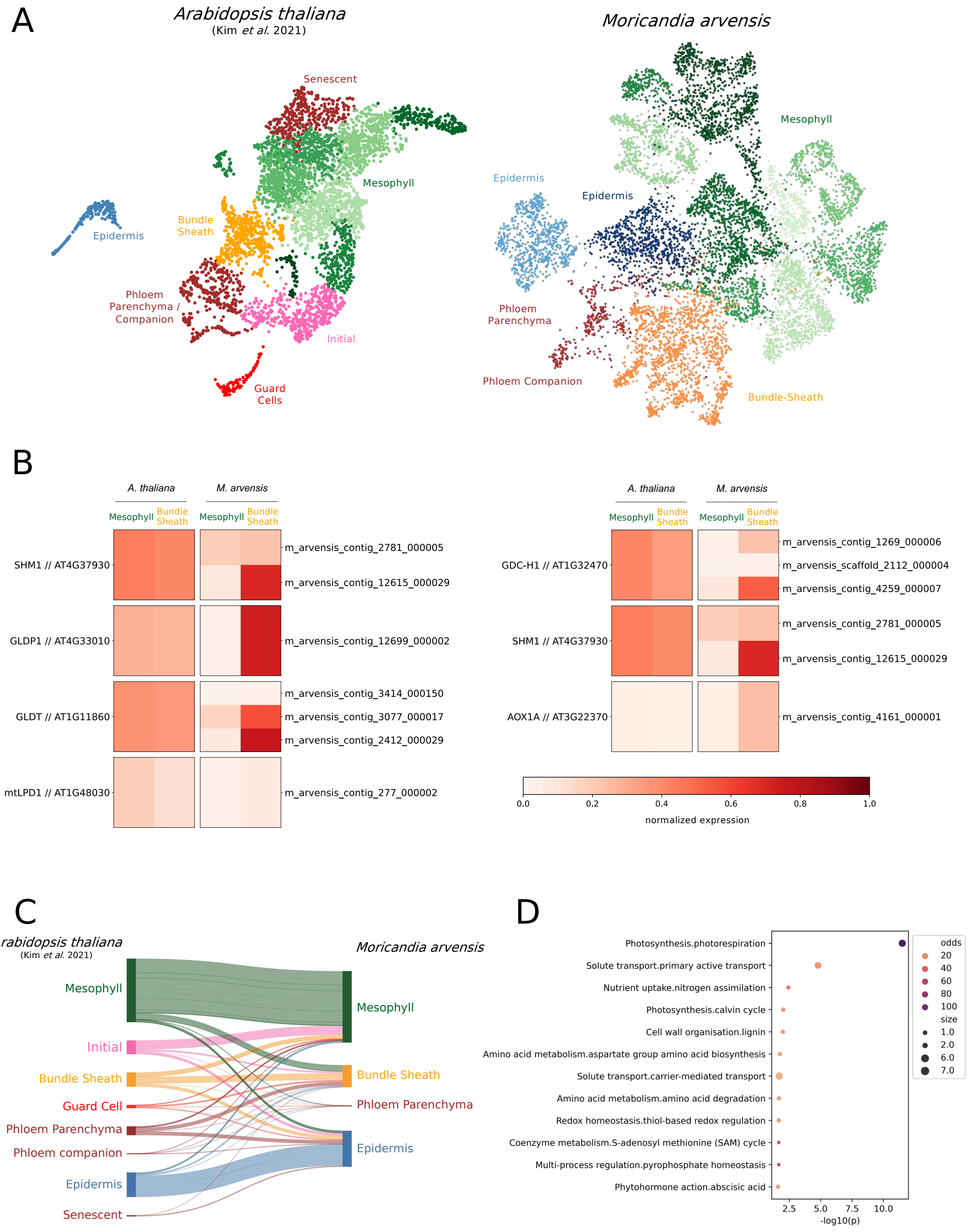


Figure 2: A: UMAP visualizations of the *A. thaliana* single-cell RNA-seq dataset from Kim *et al.* (2021; left panel) and the *M. arvensis* dataset from this study (right panel). **B: Heatmaps indicating normalized expression** in mesophyll and bundle-sheath clusters in *A. thaliana* and *M. arvensis*. Expression was normalized to the highest expressed marker gene in the respective cluster and species. **C: Sankey diagram indicating the expression of *M. arvensis* marker genes** (right side of the diagram) in *A. thaliana* clusters (left side). **D: Enrichment of *MapMan* bins** for 101 differentially partitioned genes between *M. arvensis* and *A. thaliana* bundle-sheath clusters.

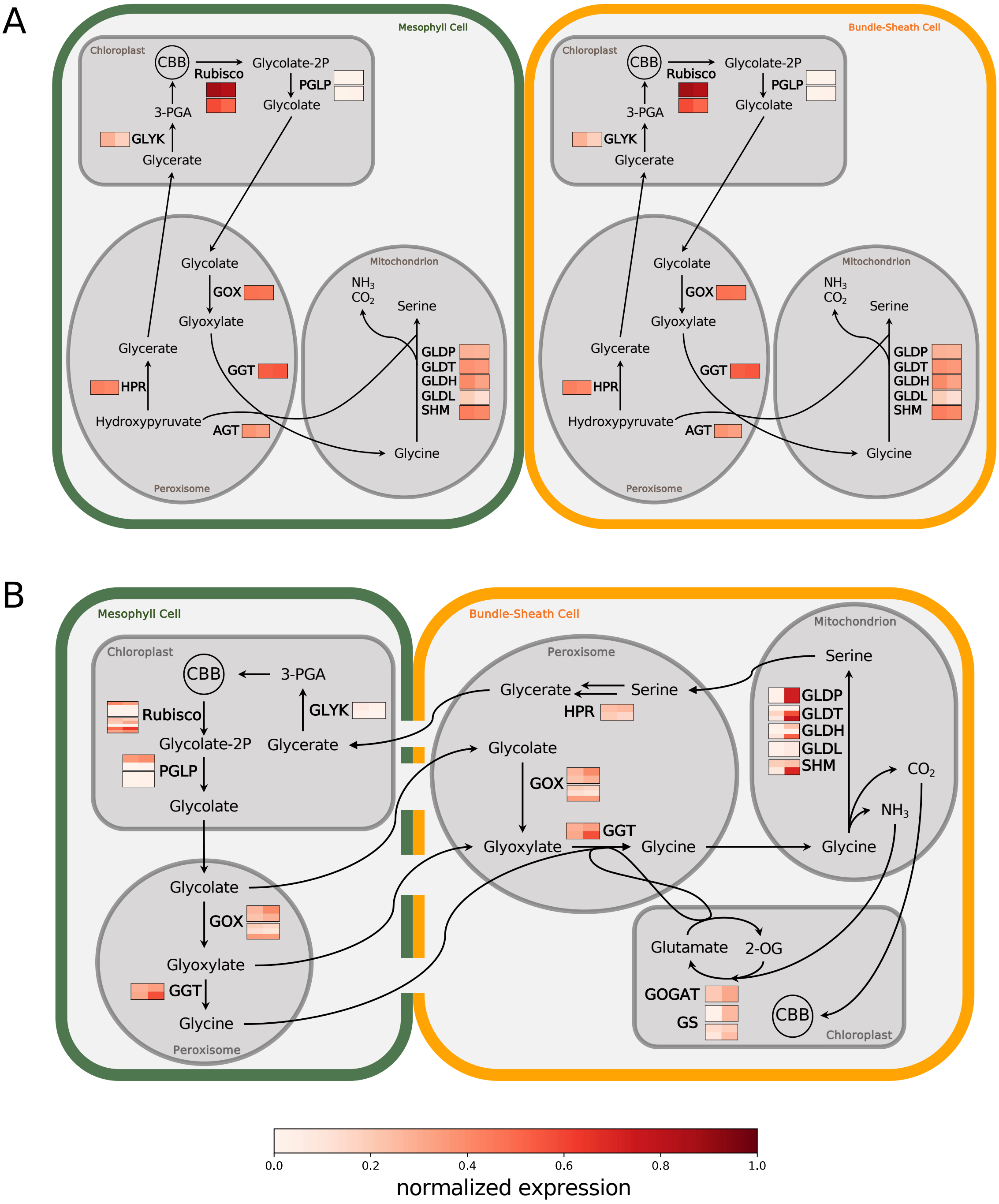


Figure 3: Schematic view of the photorespiratory pathway in *A. thaliana* (A) and *M. arvensis* (B) mesophyll and bundle sheath cells. Heatmaps indicate the expression of genes encoding selected photorespiratory enzymes from the respective *A. thaliana* and *M. arvensis* single-cell RNA-seq datasets. The expression was normalized to the highest marker genes in the respective cluster and species. The left tile of each heatmap indicates normalized expression in the mesophyll, the right tile indicates normalized expression in the bundle-sheath. Abbreviations: CBB: Calvin-Benson-Bassham Cycle, 3-PGA: 3-Phosphoglyceric acid, Rubisco: Ribulose-1,5-bisphosphate carboxylase/oxygenase, GLYK: glycerate kinase, GOX: glycolate oxidase, GGT: glutamate:glyoxylate aminotransferase, AGT: serine:glyoxylate aminotransferase, HPR: hydroxypyruvate reductase, GLDP/GLDT/GLDH/GLDL: glycine decarboxylase P/T/H/L protein, SHM: serine hydroxymethyltransferase, PGLP: phosphoglycolate phosphatase, GOGAT: glutamine-2-oxoglutarate aminotransferase, GS: glutamine synthetase, 2-OG: 2-oxoglutaric acid.

For *A. thaliana*, binding sites for various co-expressed TFs were identified in the marker gene upstream regions. One prime example was the accumulation of MYB30 TF binding sites in the *A. thaliana* epidermis marker genes, which was previously associated with epidermal wax synthesis (Zhang *et al.* (2019); Fig. 4A). Furthermore, DOF1 and DOF5 TFs were co-expressed and predicted to bind BSC marker genes in the *A. thaliana* data, which is supported by previous studies (Guo *et al.*, 2009; Dai *et al.*, 2022). This indicated that our pipeline for the identification of co-expressed TF binding sites was robust. In the *A. thaliana* dataset, a larger number of co-expressed and potentially binding TFs was identified than in the *M. arvensis* data. This was likely an effect of the lower read coverage in the *M. arvensis* snRNA-seq data.

In the *M. arvensis* BSC marker genes, two co-expressed and putatively binding TFs were identified, TGACG MOTIF-BINDING FACTOR 4 (TGA4) and NAC DOMAIN CONTAINING PROTEIN 83 (NAC083; Fig. 4A). Interestingly, these TFs were not identified as co-expressed and binding in *A. thaliana* BSC marker genes. We analyzed STARR-seq data (Jores *et al*., unpublished) for the expression strength of enhancers containing a TGA binding motif (AC(G/A)TCA) or a NAC binding motif (TT(G/A)CGT). In doing so, we found that the expression strength conveyed by these motifs in tobacco leaves decreased significantly in the light compared to the dark, indicating light-responsiveness of the genes controlled by TGA and NAC TFs (Supp. Figure S8).

In order to facilitate the analysis of *cis*-regulation underlying cell-specific expression or differential partitioning, we developed a single-cell browser for our *M. arvensis* dataset. The expression strength of each gene can be visualized using the gene name, AGI code or the *M. arvensis* gene ID as input. To visualize the available NGS data from experimental work or deep learning predictions for the locus of interest, we integrated a “multiome viewer” option to the browser. Using the multiome viewer, intron/exon structures from *Helixer* (Stiehler *et al.*, 2020; Holst *et al.*, 2023), 5-methylcytosine DNA modification intensities from whole-genome bisulfite sequencing (Chapter 9), chromatin accessibility predictions from *Predmoter* (Kindel *et al.*, 2024) and enhancer and core promoter regions identified using STARR-seq (Jores *et al.*, unpublished) can be visualized (Fig. 4B). To visualize single-cell and multiome results, a *shiny* webserver was deployed and is accessible under https://134.99.200.66:8000 from the internal network of Heinrich Heine University Düsseldorf.

A striking example for *cis*-variation underlying differential partitioning is the *GLDP1* gene encoding the GDC P-protein. Here, a large area with high levels of 5-methylcytosine DNA modification could be observed (Fig. 4B), which corresponded to the transposable element in the *GLDP1* upstream region identified in (Triesch *et al.* (2024), Chapter 8).

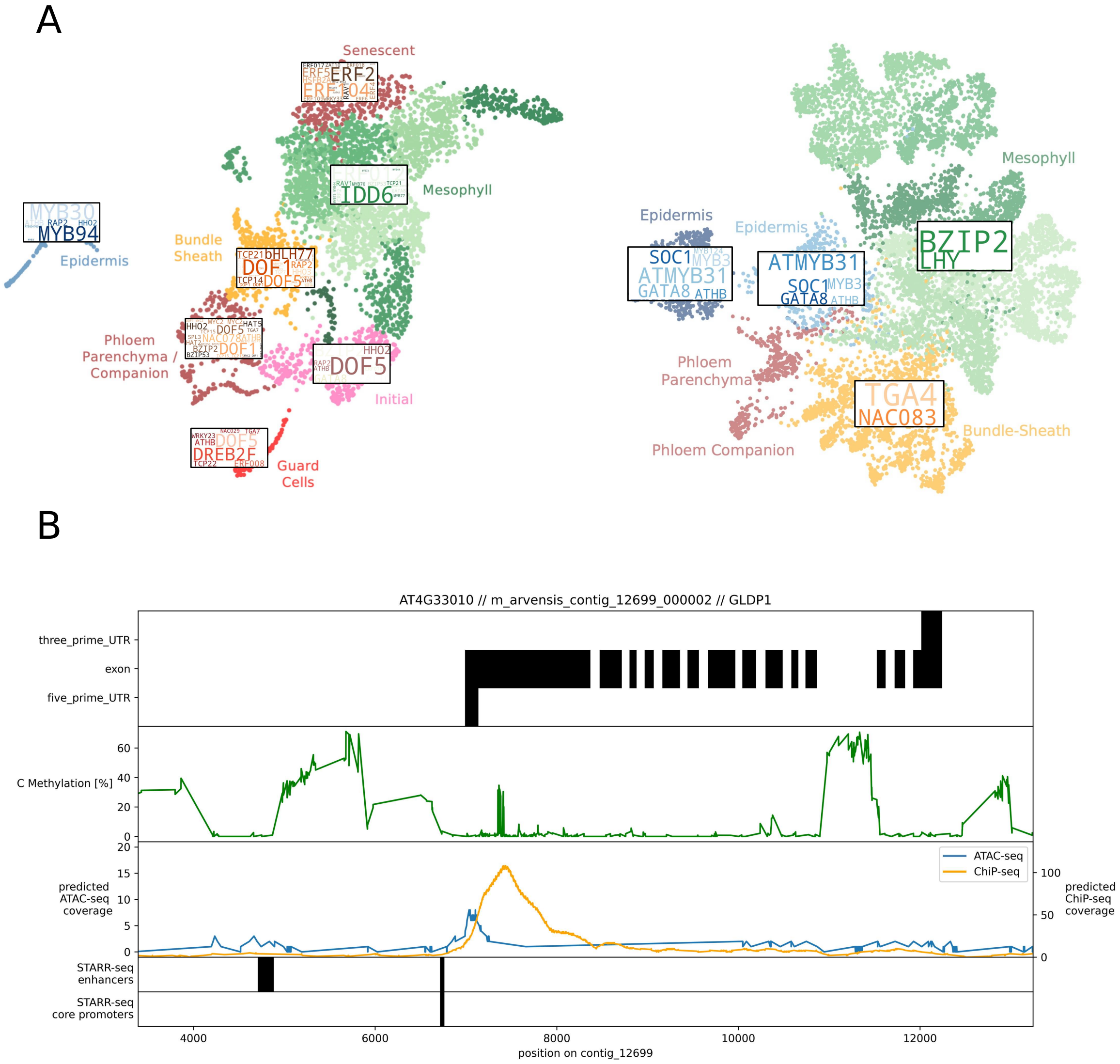


Figure 4: A: UMAP visualizations of the *A. thaliana* single-cell RNA-seq dataset from Kim *et al.* (2021; left panel) and the *M. arvensis* dataset from this study (right panel). Word clouds indicate the relative frequency of binding sites for co-expressed transcription factors in the 3000 bp upstream region of marker genes for each cluster. **B**: **Multiome summary plot for the *GLDP1* gene**. Top row: Annotated gene structure predictions from *Helixer* (Stiehler *et al.*, 2020; Holst *et al.*, 2023). Second row: 5-methylcytosine DNA methylation levels obtained using whole-genome bisulfite sequencing (Chapter 9). Third row: ATAC and ChIP-seq predictions obtained from *Predmoter* (Kindel *et al.*, 2023). Third and fourth row: STARR-seq peaks for core promoter and enhancer regions.

Discussion

While single-cell or tissue-enriched transcriptomes were generated for C_4 plants (Döring *et al.*, 2016; Liu *et al.*, 2022; Hoang *et al.*, 2023; Swift *et al.*, 2023), spatially resolved gene expression studies for a C_3-C_4 intermediate species lag significantly behind. In the Brassicaceae, the presence of multiple independent origins of C_3-C_4 intermediate photosynthesis harbors potential to map common genetic features to common traits. The whole model of the Brassicaceae C_3-C_4 intermediate metabolism relies on the observation that the P-protein of the GDC localizes specifically to BSC mitochondria (Rawsthorne *et al.*, 1988b; Rawsthorne *et al.*, 1988a) and that glycine, serine, glutamate, alanine and malate levels are significantly increased in *M. arvensis* (Schlüter *et al.*, 2016). However, the long-standing aim, introducing C_3-C_4 intermediate or even C_4 traits into C_3 plants, will require a profound knowledge about the transcriptional patterning and the functional genetics underlying these traits.

Using snRNA-seq of young *M. arvensis* leaves we provide here the *bona fide* first single-cell resolution transcriptome for a C_3-C_4 intermediate Brassicaceae species. Our dataset covers the majority of leaf cell types and shows robustness between the biological replicates (Supp. Figure S1). We assigned cell identities to clusters based on marker genes, which are consistent with previous plant leaf single-cell transcriptomes (Berrío *et al.*, 2022; Procko *et al.*, 2022; Swift *et al.*, 2023).

To compare the single-cell transcriptome data in the leaf with C_3-C_4 photosynthesis to a C_3 photosynthesis species, we used publicly available data from *A. thaliana* that was generated using leaf protoplasts (Kim *et al.*, 2021). The split between the Brassiceae and the Camelinae tribes containing *M. arvensis* and *A. thaliana*, respectively, dates back 25 million years (Walden *et al.*, 2020). However, the high number of shared marker genes between the respective cell types allows intra-species comparisons

(Suppl. Fig S3). Using the homologs for a set of *A. thaliana* marker genes, we were able to assign cell identities to the *M. arvensis* clusters with high fidelity. The use of *MMseqs2* (Steinegger *et al.*, 2017) to identify *A. thaliana* homologs for the *M. arvensis* gene models proved highly efficient, although it does not faithfully distinguish between paralogs. However, since we were interested in the differential expression or differential partitioning of paralogs, their inclusion was beneficial. Being derived from leaf protoplast, the *A. thaliana* single-cell expression dataset had a higher read count per cell compared to our *M. arvensis* dataset, which did result in the detection of more transcripts per cell. For the highly expressed photorespiratory genes, the low read count in the *M. arvensis* dataset was no obstacle, as seen in the high number of differentially partitioned genes detected.

***Moricandia arvensis* single-nuclei RNA sequencing gives insights into bundle-sheath functionalization**

Due to the prominent role of the BSC in the C_3-C_4 intermediate leaf, we analyzed differential gene expression patterns between BSC in MC in *M. arvensis* as well as between the *M. arvensis* and *A. thaliana* BSC clusters. In doing so, we noticed that multiple BSC markers genes in *M. arvensis* were not BSC-specific in *A. thaliana.* The majority of these differentially partitioned genes showed MC and phloem parenchyma expression in *A. thaliana* (Fig. 2C). The functionalization of the BSC by recruiting gene expression from MC and phloem cells was also observed in a comparative single-cell transcriptomic study between sorghum and rice (Swift *et al.*, 2023) and is thought to be one of the crucial drivers of C_4 evolution (Hibberd *et al.*, 2010; Reeves *et al.*, 2017; Singh *et al.*, 2023). This functionalization entails the shift of photorespiration to the BSC as well as an altered leaf ultrastructure with an increased amount of centripetally arranged plastids (Sage *et al.*, 2012; Heckmann *et al.*, 2013; Schlüter *et al.*, 2017).

To find the genes underlying both phenomena, we analyzed the set of differentially partitioned genes between the two species in greater detail. First, we observed a strong enrichment in photorespiratory gene expression in the *M. arvensis* BSC compared to *A. thaliana* (Fig. 1C). This was most pronounced in the differential partitioning of GDC proteins in the BSC of the C_3-C_4 intermediate leaf (Fig. 1D). Previous studies largely focused the GDC-P protein and showed BSC-specific localization of this protein, but they reported that the other subunits of the GDC were not BSC-specific (Hylton *et al.*, 1988; Morgan *et al.*, 1993). Contrasting, we were able to show that the genes encoding the T-, L- and H- protein of the GDC underly strong BSC-preferential expression in *M. arvensis*.

In our dataset, the genes encoding SHM and BOU also exhibit strong BSC-preferential expression in *M. arvensis*. In Eisenhut *et al.* (2013), it was observed that *BOU* expression correlates with the expression of genes encoding GDC proteins and *bou* mutants exhibit decreased GDC activity. The authors suggested that BOU may transport a GDC cofactor into the mitochondria. It is conceivable that BOU transports glutamate for the polyglutamylation of THF, a cosubstrate for the GDC and SHM. The differential distribution of BOU in the C_3-C_4 intermediate BSC may ensure BSC-specific GDC activity and may influence C1 metabolism. Consequently, altered C1 metabolism in the BSC could represent another layer of gene regulation through DNA methylation, which is closely tied to the C1 cycle (Groth *et al.*, 2016). Other genes encoding transporter proteins such as *UMAMIT36* and multiple ABC-transporters were also differentially partitioned in the *M. arvensis* BSC. Again, this hints at the large functionalization of this cell type and the importance of organellar and intra-cellular metabolite exchange in the C_3-C_4 intermediate leaf.

Glycolate and glyoxylate are parsimonious additions to the glycine shuttle

The photorespiratory glycine shuttle stands out as a hallmark characteristic of species exhibiting C_3-C_4 intermediate photosynthesis, sometimes referred to as C_2 species due to this trait. However, the glycine shuttle leads to a large nitrogen imbalance between MC and BSC, since two molecules of glycine are converted to one molecule of serine, releasing one net molecule of NH_3. Several hypothetical nitrogen rebalancing pathways were proposed, largely based on flux modeling (Mallmann *et al.* (2014), Supp. Fig. S6). These nitrogen shuttles resemble initial C_4 fluxes and their installation is conceived as a driver of C_4 evolution (Lundgren *et al.*, 2017).

When examining our single-cell transcript data, we realized discrepancies between the observed and proposed transcript levels for hypothetical nitrogen back-shuttle pathways. For example, we did not observe increased transcript levels for *PEP CARBOXYLASE* (*PEPC)*, *PYRUVATE ORTHOPHOSPHATE DIKINASE* (*PPDK*), *ALANINE AMINOTRANSFERASE* (*AlaAT*) and *MALATE DEHYDROGENASE* (*MDH*) or a strong expression shift of these genes to the MC (Fig. 3, Supp. Fig. S6). However, we observed an expression shift of *GLYCOLATE OXIDASE* (*GOX*), *GLUTAMATE:GLYOXYLATE AMINOTRANSFERASE* (*GGAT*), *GLUTAMINE-2-OXOGLUTARATE AMINOTRANSFERASE* (*GOGAT*) and *GLUTAMINE SYNTHETASE* (*GS*) to the BSC. Higher enzyme activities in *M. arvensis* protoplast fractions were found for GDC, GOGAT, GS, but not for HPR and GOX in the BSC compared to the MC fractions (Rawsthorne *et al.*, 1988b; Rawsthorne *et al.*, 1988a). Based on this, we hypothesize that glycolate and glyoxylate serve as additional shuttle molecules from MC to BSC, where they are converted to glycine (Fig. 3B). The shift of two additional photorespiratory conversions to the BSC would be consistent with the strong activation of the BSC in the C_3-C_4 intermediate leaf. The conversion of glyoxylate to glycine is catalyzed by GGT and dependent on glutamate and 2-oxoglutarate as co-substrates. The recycling of glutamate is catalyzed by GOGAT and GS and binds additional nitrogen, like the photorespiratory NH_3 released by the GDC. Refixation of photorespiratory NH_3 in the BSC could alleviate the nitrogen imbalance and the need for a complex nitrogen back-shuttle (Borghi *et al.*, 2022). The lack of such a complex shuttle that even resembles a primordial C_4 cycle would explain the evolutionary stability of C_3-C_4 intermediate Brassicaceae. However, since transcript abundance does not resemble protein level in a specific cell type, protein levels of the respective enzymes must be resolved using cell-specific proteomics or quantitative protein localization studies.

The photorespiratory glycine shuttle is integrated into the bundle-sheath metabolism by spatially constrained gene expression

The shift of the glycine decarboxylating steps of photorespiration to the C_3-C_4 intermediate BSC is mostly likely associated with a further rearrangement of the BSC metabolism, as shown for the sulfate and nitrogen assimilation pathways in C_3-C_4 intermediate *Flaveria* plants (Mallmann *et al.*, 2014; Weckopp *et al.*, 2015). We found a higher BSC expression of genes involved in sulfur metabolism in *A. thaliana* compared to *M. arvensis* (Supp. Fig. S5). This included especially the primary reductive steps of sulfate assimilation that were highly BSC-specific in both species. This specificity, however, was more pronounced in *A. thaliana*. Conversely, the strong BSC-preferential expression of *SULTR2;2* and *SULTR3;3*, encoding leaf sulfate transporters, was more pronounced in *M. arvensis*. This was remarkable because it confirmed the hypothesis that the confinement of sulfur assimilation to the BSC is not a prerequisite for a C_3-C_4 intermediate phenotype (Koprivova *et al.*, 2001; Kopriva *et al.*, 2005).

It was shown that nitrate reduction occurs primarily in the C_4 MC to avoid competition for reduction equivalents between sulfate and nitrate homeostasis in the BSC (Aubry *et al.*, 2014). As the shift of photorespiratory nitrogen reduction towards the BSC in *M. arvensis* is likely connected to the overall nitrogen metabolism, it might be plausible to assume that sulfur metabolism in turn loses its BSC preference. In accordance, nitrogen transporters of the NPF family were BSC-preferential in both analyzed species (Supp. Data on GitLab).

We also interrogated genes involved in starch and sucrose synthesis as well as in primary metabolism but found no difference in BSC-specificity between the two species. For example, the characteristic high expression of genes encoding SUCROSE PHOSPHATASE (SPP) enzymes was observable in both *A. thaliana* and *M. arvensis* BSC. We conclude that the overall sugar metabolism in the C_3-C_4 intermediate *Moricandia* is therefore similar to C_3 *A. thaliana*. We could not make conclusions regarding aspects of the primary carbon metabolism *M. arvensis* since multiple genes were missing in our dataset. This included genes involved in the tricarboxylic acid (TCA) cycle, glycolysis and amino acid synthesis.

It was plausible to assume that changes in the BSC metabolism between the C_3 *A. thaliana* and the C_3-C_4 intermediate *M. arvensis* require changes in redox balancing and energy homeostasis. Our dataset revealed that genes encoding proteins involved in alternative mitochondrial respiratory chains such as ALTERNATIVE OXIDASE 1 A (AOX1A) and ALTERNATIVE NAD(P)H DEHYDROGENASE 1 (NAD1) were differentially partitioned towards the *M. arvensis* BSC. An upregulation of *AOX* gene expression was already observed in bulk transcriptome studies in *M. arvensis* but not in the C_3 sister species *M. moricandioides*. The authors concluded that increased AOX protein levels indicate enhanced re-balancing of the redox metabolism due to the glycine shuttle (Schlüter *et al.*, 2017). In our data, *AOX1a* was significantly confined to the BSC, pointing to the BSC as the primary site for photorespiratory redox state re-balancing. In *Flaveria*, species, *AOX* transcript levels were observed to be increased in C_3-C_4 intermediate species, but not in C_3 or C_4 species. Schlüter *et al.* (2017)

hypothesize that the redox balance is adjusted once a full C_4 cycle is implemented, and the *AOX* transcript increase is only necessary in C_3-C_4 intermediate species. However, it needs to be verified whether the BSC specificity is also reverted in C_4 species. Furthermore, we cannot rule out that BSC-specific expression of *AOX* is also found in closely related C_3 *Moricandia* species. Redox homeostasis can also be maintained by the malate valve, especially through the interplay of NADP-MDH (NADP-dependent malate dehydrogenase) enzymes (Selinski *et al.*, 2019). Whereas the plastidial NADP-MDH is encoded by a single gene in *A. thaliana*, it is expressed by three paralogs in *M. arvensis*. It was interesting to find a differential partitioning between these paralogs in *M. arvensis*. This could hint at a tight control of the malate metabolism between MC and BSC, carried out by multiple genes with potentially individual regulatory mechanisms.

We specifically analyzed the dataset with regard to genes that suggest a selectively increased or reduced connectivity between MC and BSC tissue in *M. arvensis.* We found genes associated with the formation of aquaporins, water transmembrane transporters, partitioned to the BSC in both analyzed species, but more constrained to the *A. thaliana* BSC. It was hypothesized that in C_4 plants, the elevated expression of aquaporins aids in CO_2 transportation to the mesophyll cell, increasing mesophyll conductance and by this the photosynthetic efficiency (Weber *et al.*, 2010; Kaldenhoff, 2012; Ermakova *et al.*, 2021). In the *M. arvensis* BSC, we hypothesize that the less pronounced differential expression of aquaporin-forming genes acts as an insulator to prevent exchange of oxygen and carbon dioxide between MC and BSC.

The C_3-C_4 intermediate biochemistry in *M. arvensis* leaves is accompanied by a set of ultrastructural modifications in leaf architecture. We sought to underpin these anatomical patterns with variation in single-cell transcription patterns of anatomical regulator genes. To this end, we analyzed selected leaf anatomy-related genes that were previously analyzed in bulk RNA-seq approaches in *M. arvensis* (Kong *et al*., 2011; Lin, 2020*)*. We found no significant increase in normalized expression of these genes in either *M. arvensis* or *A. thaliana*. There was also no clear shift of these genes to either MC or BSC, but a slight tendency towards higher expression of these genes in the *M. arvensis* MC (Supp. Fig. S7). This was also true for proposed master regulators of development like the *GOLDEN2* gene, where we detected a preferential expression in both *A. thaliana* and *M. arvensis* BSC (Supp. Tab. S7).

In contrast, we also found genes like AT1G67480 encoding a galactose oxidase/kelch repeat superfamily protein and AT1G64340 encoding a serine/threonine-kinase to be strongly differentially partitioned. These genes are no further classified and apparently no research was done on these genes to date. We cannot rule out that these transcripts were mapped to false gene models in our reference transcript that was generated using *in silico* predictions with *Helixer* (Stiehler *et al*., 2020; Holst *et al*., 2023). However, it is striking that these transcripts prevail in high expression strength in the BSC. It is believed that the evolution of C_4 photosynthesis is rather driven by the incorporation of existing genes into new gene regulatory networks (Williams *et al.*, 2012; Singh *et al.*, 2023; Swift *et al.*, 2023). However, the presence of multiple uncharacterized genes in the top hits of differentially partitioned transcripts in *M. arvensis* leaves room for speculation about the contribution of unknown factors in gene regulation, development or biochemistry of the C_3-C_4 intermediate species.

C_3-C_4 plants are stable intermediates on the evolutionary progression towards C_4 photosynthesis

The lack of C_4 species in the Brassicaceae family raises questions about the evolutionary state of C_3-C_4 intermediate Brassicaceae and the selective pressure towards C_4 photosynthesis. It has been discussed that the retention of the species in cooler climates or anatomical constraints decreased selection for C_4 traits (Lundgren *et al.*, 2017; Schlüter *et al.*, 2017; Lundgren, 2020; Walsh *et al.*, 2023). Here, we present multiple lines of evidence that support the hypothesis that the traits observed in C_3-C_4 intermediate Brassicaceae are distinguished phenotypes following the trajectory of C_4 evolution: Firstly, the complete shift of GDC activity predicted to be at the onset of C_4 evolution is clearly observable in our *M. arvensis* dataset (Sage *et al.*, 2012; Heckmann *et al.*, 2013). The additional flux of glycolate and glyoxylate and the BSC-specific refixation of photorespiratory NH_3 could be one of the constraints decreasing the selection of a fully evolved C_4 cycle in the Brassicaceae.

Secondly, it is widely accepted that the evolution of C_4 photosynthesis necessitates the integration of existing gene regulatory networks, such as the incorporation of C_3 light-responsiveness elements into C_4 promoters (Hibberd *et al.*, 2010; Reeves *et al.*, 2017; Singh *et al.*, 2023; Swift *et al.*, 2023). In line with this, the accumulation of light-responsive upstream elements in the *M. arvensis* BSC marker genes provides additional evidence indicating the evolutionary progression of C_3-C_4 intermediate Brassicaceae towards C_4 photosynthesis (Supp. Fig. S8). The enrichment of TGA TF binding sites in the upstream regions of *M. arvensis* BSC marker genes was especially striking, since members of this class were found to be associated with the increase in transcripts of C_4 genes during the de-etiolation of seedling of the C_4 species *Gynandropsis gynandra* (Singh *et al.*, 2023).

Finally, details about the ontogenetic development of the C_3-C_4 intermediate photosynthesis in Brassicaceae remain elusive. Using single-cell RNA velocity modelling, we tried to provide insights on the expression patterns of individual genes in various cell states (La Manno *et al.*, 2018). In doing so, we observed a strong decrease in *GLDP1* expression in cells from the *M. arvensis* MC population following the cell's modelled pseudo-age. This would indicate a high *GLDP1* expression and, thus, a C_3-like photosynthesis in young *M. arvensis* MC. However, this data is highly hypothetical and requires thorough experimental corroboration.

The advent of single-cell RNA sequencing cleared a substantial roadblock in the way of a thorough understanding of multiple complex traits (Cervantes-Pérez *et al.*, 2022; Nobori *et al.*, 2023). Using a snRNA sequencing dataset from *M. arvensis*, we were able to confirm and expand old hypothesis about the localization of photorespiratory enzymes in C_3-C_4 intermediate Brassicacea (Hylton *et al.*, 1988; Rawsthorne *et al.*, 1988b; Rawsthorne *et al.*, 1988a; Rawsthorne, 1992; Morgan *et al.*, 1993; Monson *et al.*, 2000), the functional activation of BSC tissue (Sage, 2004; Swift *et al.*, 2023) and present alternatives to proposed nitrogen shuttle pathways (Mallmann *et al.*, 2014). In addition, we gathered first insights into the recruitment of light-responsive *cis*-elements in BSC marker genes. These insights into the genetic mechanisms of C_3-C_4 intermediate photosynthesis in *M. arvensis* will help to identify targets for synthetic biology strategies to engineer C_4 traits into crop plants (Ermakova *et al.*, 2020).

Supplementary Material

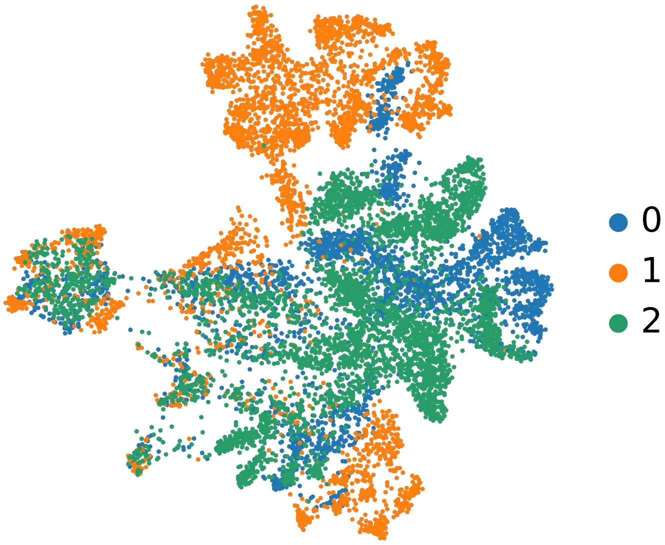


Figure S1: UMAP visualization of the *Moricandia arvensis* single-nuclei RNA-seq dataset indicating which cell belonged to which biological replicate. The number 0-2 refer to the three biological replicates.

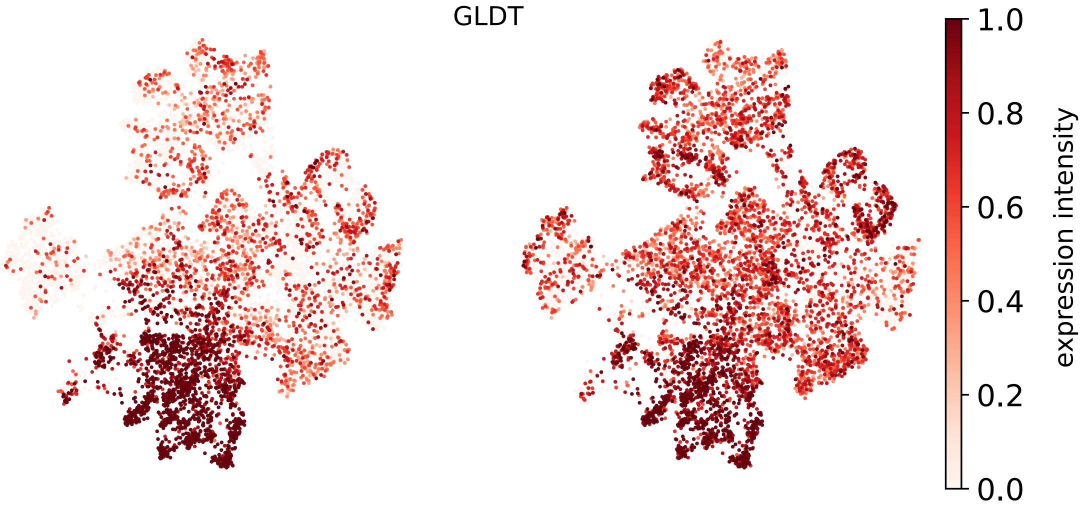


Figure S2: UMAP visualization of the *Moricandia arvensis* single-nuclei RNA-seq dataset heatmap overlays for the two *GLDT1* gene copies in *M. arvensis*.

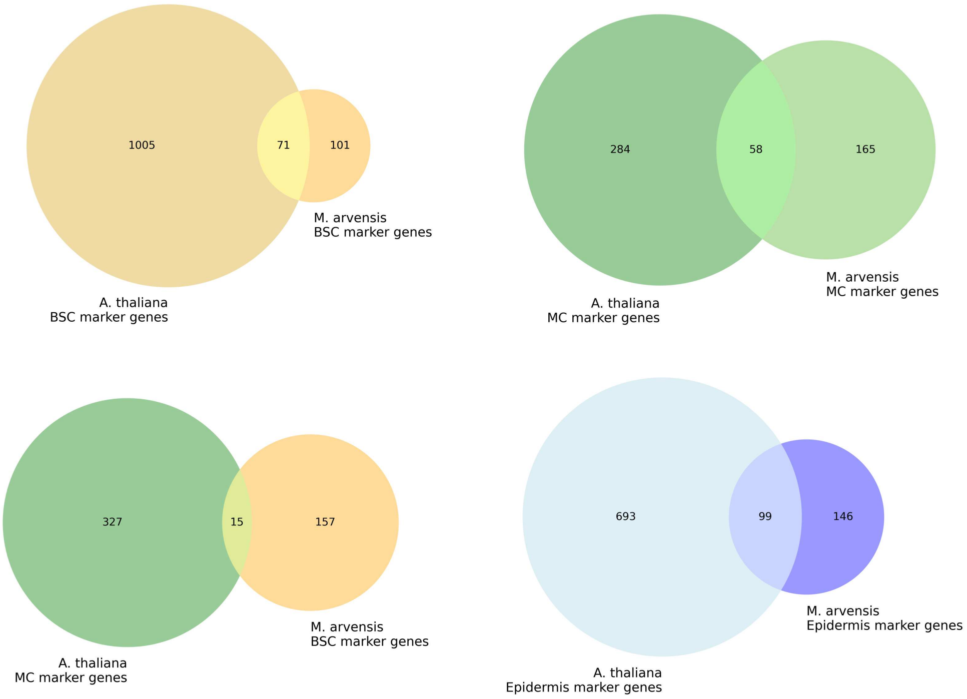


Figure S3: Venn diagram indicating the number of shared and unique marker genes in different clusters in the *Arabidopsis thaliana* and *Moricandia arvensis* leaf snRNA-seq dataset.

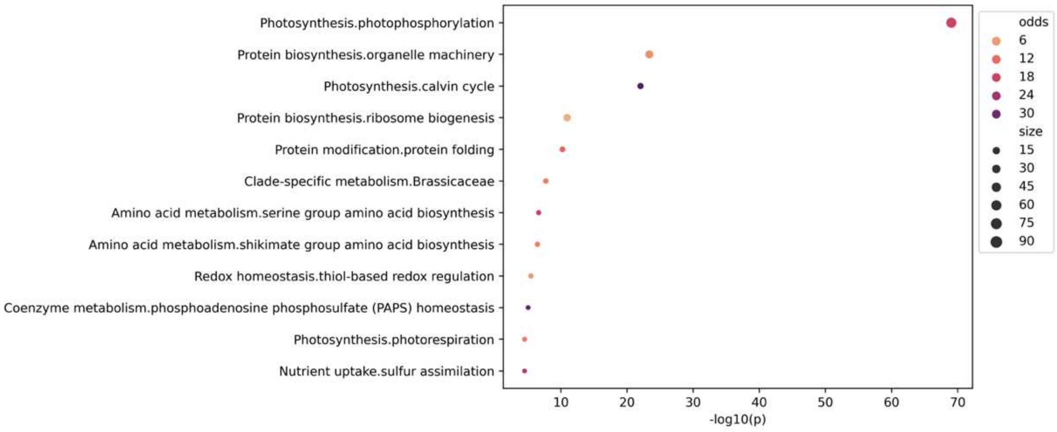
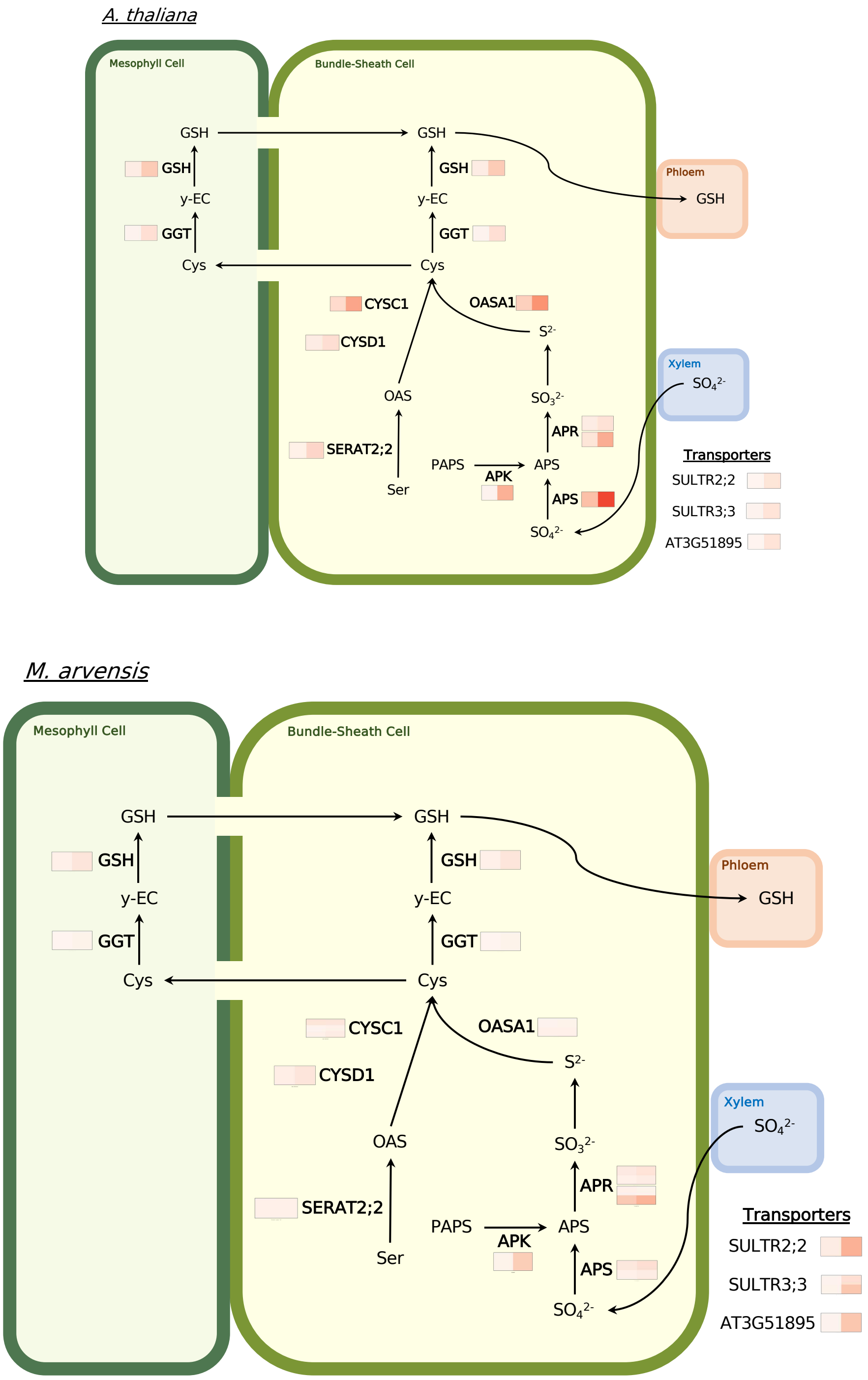


Figure S4: Enrichment of *MapMan* bins for differentially expressed genes between *Arabidopsis thaliana* bundle-sheath and mesophyll clusters.



A. thaliana

M. arvensis

Figure S5: Schematic illustration of the sulfur assimilation pathway in bundle-sheath and mesophyll cells in *Arabidopsis thaliana* (upper panel) and *Moricandia arvensis* (lower panel). The left tile of each heatmap indicates normalized expression in the mesophyll, the right tile indicates normalized expression in the bundle-sheath. Abbreviations: Ser: serine, APS: Adenosine-5'-phosphosulfate, APS (enzyme): APS sulfurylase, APR: AOS reductase, PAPS: 3′-Phosphoadenosin-5′-phosphosulfate, OAS: O-acetylserine, Cys: cysteine, y-EC: y-glutamylcysteine, GSH: glutathione, GGT: γ-glutamyltransferase, GSH (enzyme): glutamate-cysteine ligase, SULTR: sulfate transporter, CYSC/CYSD: cysteine synthase

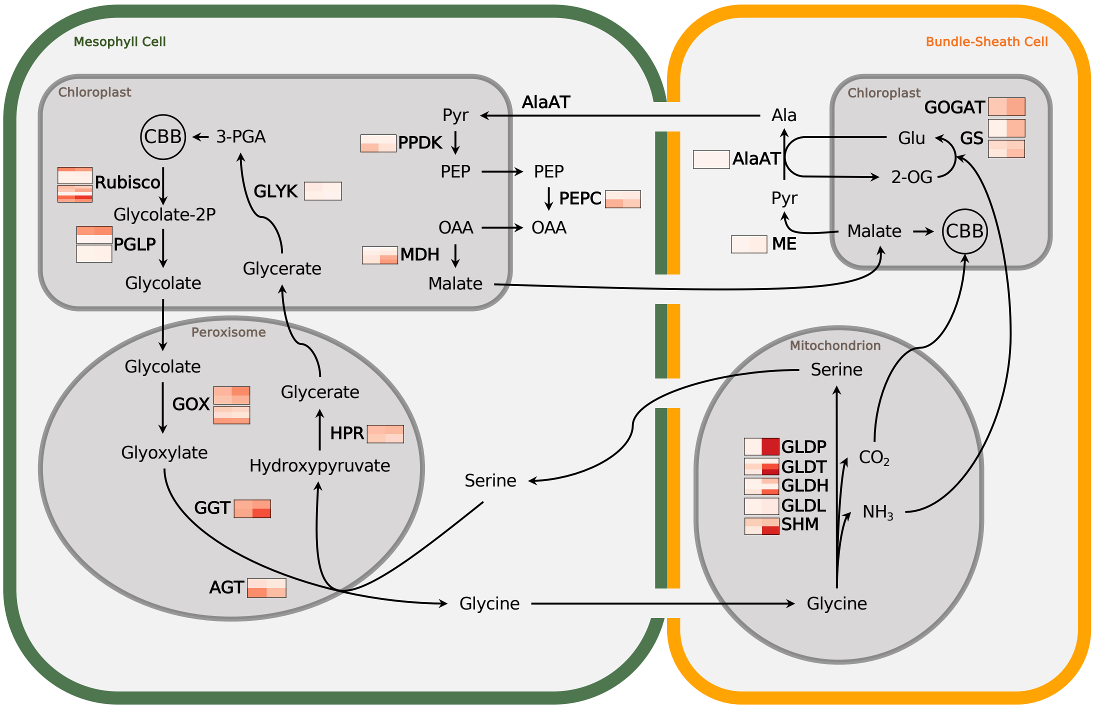


Figure S6: Schematic illustration of the photorespiratory pathway in *Moricandia arvensis* with a putative alanine-pyruvate shuttle as proposed in Mallmann *et al.* (2014). Heatmaps indicate the expression of genes encoding selected photorespiratory enzymes from the *M. arvensis* single-cell RNA-seq dataset. The expression was normalized to the highest marker genes in the respective cluster. The left tile of each heatmap indicates normalized expression in the mesophyll, the right tile indicates normalized expression in the bundle-sheath. Abbreviations: CBB: Calvin-Benson-Bassham Cycle, 3-PGA: 3-Phosphoglyceric acid, Rubisco: Ribulose-1,5-bisphosphate carboxylase/oxygenase, GLYK: glycerate kinase, GOX: glycolate oxidase, GGT: glutamate:glyoxylate aminotransferase, AGT: serine:glyoxylate aminotransferase, HPR: hydroxypyruvate reductase, GLDP/GLDT/GLDH/GLDL: glycine decarboxylase P/T/H/L protein, SHM: serine hydroxymethyltransferase, PGLP: phosphoglycolate phosphatase, GOGAT: glutamine-2-oxoglutarate aminotransferase, GS: glutamine synthetase, 2-OG: 2-oxoglutaric acid, Ala: Alanine, Pyr: Pyruvate, AlaAT: alanine aminotransferase, ME: malic enzyme, PEP: phosphoenolpyruvate, OAA: oxaloacetic acid.

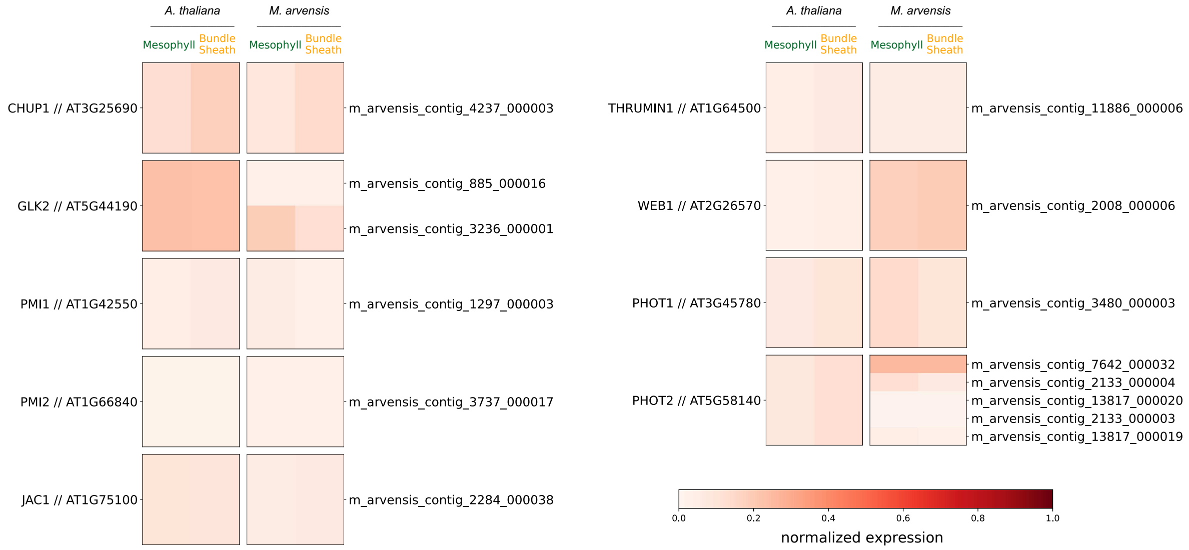


Figure S7: Heatmaps indicating normalized expression of selected anatomical regulator genes in mesophyll and bundle-sheath clusters in *Arabidopsis thaliana* and *Moricandia arvensis*. Expression was normalized to the highest expressed marker gene in the respective cluster and species

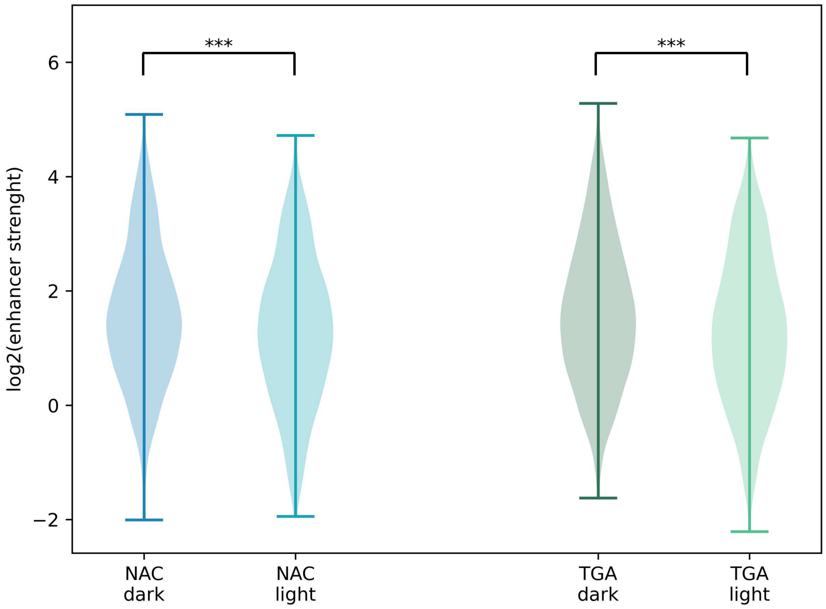


Figure S8: Violin plots showing the enhancer strength of enhancer sequences with a NAC or TGA transcription factor binding motif under dark and under light. Enhancer strength was determined using self-transcribing active regulatory region sequencing (STARR-seq) data provided by Jores *et al.* (personal communication).

Table S1: List of the top 50 differentially partitioned genes between *Moricandia arvensis* and *Arabidopsis thaliana* bundle-sheath clusters.

Gene Name	AGI	MapMan bin
GLDT	AT1G11860	Photosynthesis.photorespiration.glycine decarboxylase complex.aminomethyltransferase component *(T-protein)
GLDP1	AT4G33010	Photosynthesis.photorespiration.glycine decarboxylase complex.glycine dehydrogenase component *(P-protein)
SHM1	AT4G37930	Photosynthesis.photorespiration.serine hydroxymethyltransferase *(SHM)
AT1G67480	AT1G67480	not assigned.annotated
AT1G64340	AT1G64340	not assigned.not annotated
GDC-H1	AT1G32470	Photosynthesis.photorespiration.glycine decarboxylase complex.lipoamide-containing component *(H-protein)
BOU	AT5G46800	Solute transport.carrier-mediated transport.solute transporter *(MTCC)
ALKBH10B	AT4G02940	RNA processing.RNA modification.RNA methylation.mRNA methylation.N6-methyladenosine demethylation.demethylase *(ALKBH10)
GLDT	AT1G11860	Photosynthesis.photorespiration.glycine decarboxylase complex.aminomethyltransferase component *(T-protein)
PIP2A	AT3G53420	Solute transport.channels.MIP family.plasma membrane intrinsic protein *(PIP)
NAIP2	AT1G16520	Cell division.cytokinesis.endoplasmic reticulum (ER) reorganisation.ER body formation factor *(NAIP)
ABCC1	AT1G30400	Solute transport.primary active transport.ABC superfamily.ABC1 family.subfamily ABCC transporter
AT1G69523	AT1G69523	not assigned.not annotated
NADP-MDH	AT5G58330	Photosynthesis.calvin cycle.NADPH-dependent malate dehydrogenase *(NADP-MDH)
AT1G62480	AT1G62480	not assigned.not annotated
NMT3	AT1G73600	Lipid metabolism.glycerolipid metabolism.phosphatidylcholine biosynthesis.phospho-base N-methyltransferase
AT5G16030	AT5G16030	not assigned.not annotated
GGT1	AT1G23310	Photosynthesis.photorespiration.aminotransferase activities.glutamate-glyoxylate transaminase
LEA3	AT1G02820	not assigned.annotated
TAR4	AT1G34060	not assigned.annotated
GDC-H1	AT1G32470	Photosynthesis.photorespiration.glycine decarboxylase complex.lipoamide-containing component *(H-protein)
DELTA-TIP	AT3G16240	Solute transport.channels.MIP family.tonoplast intrinsic protein *(TIP)
AOX1A	AT3G22370	Cellular respiration.oxidative phosphorylation.alternative NAD(P)H dehydrogenase activities.alternative oxidase *(AOx)
UMAMIT36	AT1G70260	Solute transport.carrier-mediated transport.DMT superfamily.solute transporter *(UmamiT)

AOAT2	AT1G70580	Photosynthesis.photorespiration.aminotransferase activities.glutamate-glyoxylate transaminase
NDA1	AT1G07180	Cellular respiration.oxidative phosphorylation.alternative NAD(P)H dehydrogenase activities.type-II NAD(P)H dehydrogenase activities.NAD(P)H dehydrogenase *(NDA)
NET1D	AT1G03080	Cytoskeleton organisation.microfilament network.actin-membrane compartment interaction.NET-type actin-membrane nexus protein families.actin-binding protein *(NET1)
BRON	AT1G75710	not assigned.not annotated
AT5G16030	AT5G16030	not assigned.not annotated
GS2	AT5G35630	Nutrient uptake.nitrogen assimilation.ammonium assimilation.glutamine synthetase activities.plastidial glutamine synthetase *(GLN2)
SGR6	AT2G36810	not assigned.annotated
TIP2	AT3G26520	Solute transport.channels.MIP family.tonoplast intrinsic protein *(TIP)
GASA1	AT1G75750	Phytohormone action.signalling peptides.CRP (cysteine-rich-peptide) category.GASA/GAST-peptide activity.GASA-precursor polypeptide
ABCC12	AT1G30410	Solute transport.primary active transport.ABC superfamily.ABC1 family.subfamily ABCC transporter
PnsB4	AT1G18730	Photosynthesis.photophosphorylation.chlororespiration.NADH dehydrogenase-like (NDH) complex.subcomplex B.component *(PnsB4/NDF6)
GAUT15	AT3G58790	not assigned.annotated
XTH28	AT1G14720	Enzyme classification.EC_2 transferases.EC_2.4 glycosyltransferase
ALA9	AT1G68710	Lipid metabolism.lipid trafficking.endoplasmic reticulum-plastid lipid transfer.phospholipid transverse translocation.ATP-dependent machinery.ALA-ALIS flippase complex.P4-type ATPase component *(ALA)
TOL5	AT5G63640	Vesicle trafficking.endocytic trafficking.ubiquitylated cargo adaptation.ubiquitin adaptor protein *(TOL)
AT1G61800	AT1G61800	Solute transport.carrier-mediated transport.DMT superfamily.NST-TPT group.phosphometabolite transporter *(TPT/PPT/GPT/XPT)
HKT1	AT4G10310	Solute transport.carrier-mediated transport.potassium/sodium cation transporter *(HKT)
UMAMIT36	AT1G70260	Solute transport.carrier-mediated transport.DMT superfamily.solute transporter *(UmamiT)
AtHMP52	AT5G50740	not assigned.annotated
ABCC5	AT1G04120	Solute transport.primary active transport.ABC superfamily.ABC1 family.subfamily ABCC transporter
AT5G16030	AT5G16030	not assigned.not annotated
ADS2	AT2G31360	Lipid metabolism.fatty acid metabolism.fatty acid desaturation.first desaturation.acyl-CoA desaturase *(ADS/FAD5)
AT3G04450	AT3G04450	RNA biosynthesis.transcriptional regulation.GARP transcription factor superfamily.subgroup PHL transcription factor
ABCB19	AT3G28860	Phytohormone action.auxin.transport.auxin efflux transporter *(ABCB19)

Literature

Aubry, S., Smith-Unna, R. D., Boursnell, C. M., Kopriva, S. and Hibberd, J. M. (2014) 'Transcript residency on ribosomes reveals a key role for the Arabidopsis thaliana bundle sheath in sulfur and glucosinolate metabolism', *Plant Journal*, 78(4), pp. 659-673.

Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., Ren, J., Li, W. W. and Noble, W. S. (2009) 'MEME Suite: Tools for motif discovery and searching', *Nucleic Acids Research*.

Bauwe, H., Hagemann, M. and Fernie, A. R. (2010) Photorespiration: players, partners and origin. *Trends in Plant Science.* Elsevier Current Trends.

Berardini, T. Z., Reiser, L., Li, D., Mezheritsky, Y., Muller, R., Strait, E. and Huala, E. (2015) 'The arabidopsis information resource: Making and mining the "gold standard" annotated reference plant genome', *genesis*, 53(8), pp. 474-485.

Berrío, R. T., Verstaen, K., Vandamme, N., Pevernagie, J., Achon, I., van Duyse, J., van Isterdael, G., Saeys, Y., de Veylder, L., Inzé, D. and Dubois, M. (2022) 'Single-cell transcriptomics sheds light on the identity and metabolism of developing leaf cells', *Plant Physiology*, 188(2), pp. 898-918.

Borba, A. R., Reyna-Llorens, I., Dickinson, P. J., Steed, G., Gouveia, P., Górska, A. M., Gomes, C., Kromdijk, J., Webb, A. A. R., Saibo, N. J. M. and Hibberd, J. M. (2023) 'Compartmentation of photosynthesis gene expression in C4 maize depends on time of day', *Plant Physiology*, 193(4), pp. 2306-2320.

Borghi, G. L., Arrivault, S., Günther, M., Barbosa Medeiros, D., Dell'Aversana, E., Fusco, G. M., Carillo, P., Ludwig, M., Fernie, A. R., Lunn, J. E. and Stitt, M. (2022) 'Metabolic profiles in C3, C3-C4intermediate, C4-like, and C4species in the genus Flaveria', *Journal of Experimental Botany*, 73(5), pp. 1581-1601.

Burgess, S. J., Reyna-Llorens, I., Stevenson, S. R., Singh, P., Jaeger, K. and Hibberd, J. M. (2019) 'Genome-Wide Transcription Factor Binding in Leaves from C3 and C4 Grasses', *The Plant Cell*, 31(10), pp. 2297-2314.

Castro-Mondragon, J. A., Riudavets-Puig, R., Rauluseviciute, I., Berhanu Lemma, R., Turchi, L., Blanc-Mathieu, R., Lucas, J., Boddie, P., Khan, A., Perez, N. M., Fornes, O., Leung, T. Y., Aguirre, A., Hammal, F., Schmelter, D., Baranasic, D., Ballester, B., Sandelin, A., Lenhard, B., Vandepoele, K., Wasserman, W. W., Parcy, F. and Mathelier, A. (2022) 'JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles', *Nucleic Acids Research*, 50(D1), pp. D165-D173.

Cervantes-Pérez, S. A., Thibivillliers, S., Tennant, S. and Libault, M. (2022) 'Review: Challenges and perspectives in applying single nuclei RNA-seq technology in plant biology', *Plant science : an international journal of experimental plant biology*, 325.

Chen, H., Yin, X., Guo, L., Yao, J., Ding, Y., Xu, X., Liu, L., Zhu, Q.-H., Chu, Q. and Fan, L. (2021) 'PlantscRNAdb: A database for plant single-cell RNA analysis'.

Dai, X., Tu, X., Du, B., Dong, P., Sun, S., Wang, X., Sun, J., Li, G., Lu, T., Zhong, S. and Li, P. (2022) 'Chromatin and regulatory differentiation between bundle sheath and mesophyll cells in maize', *Plant Journal*, 109(3), pp. 675-692.

Deal, R. B., Henikoff, S., Division, S. and Hutchinson, F. (2020) 'The INTACT method for cell type-specific gene expression and chromatin profiling in Arabidopsis', 6(1), pp. 56-68.

Döring, F., Streubel, M., Bräutigam, A. and Gowik, U. (2016) 'Most photorespiratory genes are preferentially expressed in the bundle sheath cells of the C4 grass Sorghum bicolor', *Journal of Experimental Botany*, 67(10), pp. 3053-3064.

Eisenhut, M., Planchais, S., Cabassa, C., Guivarch, A., Justin, A. M., Taconnat, L., Renou, J. P., Linka, M., Gagneul, D., Timm, S., Bauwe, H., Carol, P. and Weber, A. P. M. (2013) 'Arabidopsis A BOUT DE SOUFFLE is a putative mitochondrial transporter involved in photorespiratory metabolism and is required for meristem growth at ambient CO_2 levels', *The Plant journal : for cell and molecular biology,* 73(5), pp. 836-849.

Ermakova, M., Arrivault, S., Giuliani, R., Danila, F., Alonso-Cantabrana, H., Vlad, D., Ishihara, H., Feil, R., Guenther, M., Borghi, G. L., Covshoff, S., Ludwig, M., Cousins, A. B., Langdale, J. A., Kelly, S., Lunn, J. E., Stitt, M., von Caemmerer, S. and Furbank, R. T. (2020) 'Installation of C4 photosynthetic pathway enzymes in rice using a single construct', *Plant Biotechnology Journal*.

Ermakova, M., Osborn, H., Groszmann, M., Bala, S., Bowerman, A., McGaughey, S., Byrt, C., Alonso-Cantabrana, H., Tyerman, S., Furbank, R. T., Sharwood, R. E. and von Caemmerer, S. (2021) 'Expression of a CO2-permeable aquaporin enhances mesophyll conductance in the C4 species setaria viridis', *eLife,* 10.

Giacomello, S. (2021) 'A new era for plant science: spatial single-cell transcriptomics', *Current Opinion in Plant Biology,* 60, pp. 102041-102041.

Grant, C. E., Bailey, T. L. and Noble, W. S. (2011) 'FIMO: scanning for occurrences of a given motif', *Bioinformatics,* 27(7), pp. 1017-1018.

Groth, M., Moissiard, G., Wirtz, M., Wang, H., Garcia-Salinas, C., Ramos-Parra, P. A., Bischof, S., Feng, S., Cokus, S. J., John, A., Smith, D. C., Zhai, J., Hale, C. J., Long, J. A., Hell, R., Díaz De La Garza, R. I. and Jacobsen, S. E. (2016) 'MTHFD1 controls DNA methylation in Arabidopsis', *Nature Communications 2016 7:1,* 7(1), pp. 1-13.

Guerreiro, R., Bonthala, V. S., Schlüter, U., Hoang, N. V., Triesch, S., Schranz, M. E., Weber, A. P. M. and Stich, B. (2023) 'A genomic panel for studying C3-C4 intermediate photosynthesis in the Brassiceae tribe', *Plant Cell and Environment,* 46(11), pp. 3611-3627.

Guo, Y., Qin, G., Gu, H. and Qu, L. J. (2009) 'Dof5.6/HCA2, a Dof Transcription Factor Gene, Regulates Interfascicular Cambium Formation and Vascular Tissue Development in Arabidopsis', *The Plant Cell,* 21(11), pp. 3518-3534.

Heckmann, D., Schulze, S., Denton, A., Gowik, U., Westhoff, P., Weber, A. P. M. and Lercher, M. J. (2013) 'Predicting C4 photosynthesis evolution: Modular, individually adaptive steps on a mount fuji fitness landscape', *Cell,* 153(7), pp. 1579-1579.

Hibberd, J. M. and Covshoff, S. (2010) 'The regulation of gene expression required for C4 photosynthesis', *Annual Review of Plant Biology,* 61, pp. 181-207.

Hoang, N. V., Sogbohossou, E. O. D., Xiong, W., Simpson, C. J. C., Singh, P., Walden, N., Van Den Bergh, E., Becker, F. F. M., Li, Z., Zhu, X. G., Brautigam, A., Weber, A. P. M., Van Haarst, J. C., Schijlen, E. G. W. M., Hendre, P. S., Van Deynze, A., Achigan-Dako, E. G., Hibberd, J. M. and Schranz, M. E. (2023) 'The Gynandropsis gynandra genome provides insights into whole-genome duplications and the evolution of C4 photosynthesis in Cleomaceae', *The Plant cell,* 35(5), pp. 1334-1359.

Holst, F., Bolger, A., Günther, C., Maß, J., Triesch, S., Kindel, F., Kiel, N., Saadat, N., Ebenhöh, O., Usadel, B., Schwacke, R., Bolger, M., Weber, A. P. M. and Denton, A. K. (2023) 'Helixer–de novo Prediction of primary eukaryotic gene models conbining deep learning and a Hidden Marcov Model', *bioRxiv*.

Hua, L., Stevenson, S. R., Reyna-Llorens, I., Xiong, H., Kopriva, S. and Hibberd, J. M. (2021) 'The bundle sheath of rice is conditioned to play an active role in water transport as well as sulfur assimilation and jasmonic acid synthesis', *Plant Journal,* 107(1), pp. 268-286.

Hylton, C. M., Rawsthorne, S., Smith, A. M., Jones, D. A. and Woolhouse, H. W. (1988) 'Glycine decarboxylase is confined to the bundle-sheath cells of leaves of C3-C4 intermediate species', *Planta,* 175(4), pp. 452-459.

Kaldenhoff, R. (2012) 'Mechanisms underlying CO2 diffusion in leaves', *Current Opinion in Plant Biology,* 15(3), pp. 276-281.

Kim, J. Y., Symeonidi, E., Pang, T. Y., Denyer, T., Weidauer, D., Bezrutczyk, M., Miras, M., Zöllner, N., Hartwig, T., Wudick, M. M., Lercher, M., Chen, L. Q., Timmermans, M. C. P. and Frommer, W. B. (2021) 'Distinct identities of leaf phloem cells revealed by single cell transcriptomics', *Plant Cell,* 33(3), pp. 511-530.

Kindel, F., Triesch, S., Schlüter, U., Randarevitch, L. A., Reichel-Deland, V., Weber, A. P. M. and Denton, A. K. (2024) 'Predmoter-cross-species prediction of plant promoter and enhancer regions'.

Kong, S. G. and Wada, M. (2011) 'New Insights into Dynamic Actin-Based Chloroplast Photorelocation Movement', *Molecular Plant,* 4(5), pp. 771-781.

Kopriva, S. and Koprivova, A. (2005) 'Sulfate assimilation and glutathione synthesis in C4 plants', *Photosynthesis Research,* 86(3), pp. 363-372.

Koprivova, A., Melzer, M., Von Ballmoos, P., Mandel, T., Brunold, C. and Kopriva, S. (2001) 'Assimilatory Sulfate Reduction in C3, C3-C4, and C4 Species of Flaveria', *Plant Physiology,* 127(2), pp. 543-550.

La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M. E., Lönnerberg, P., Furlan, A., Fan, J., Borm, L. E., Liu, Z., van Bruggen, D., Guo, J., He, X., Barker, R., Sundström, E., Castelo-Branco, G., Cramer, P., Adameyko, I., Linnarsson, S. and Kharchenko, P. V. (2018) 'RNA velocity of single cells', *Nature 2018 560:7719,* 560(7719), pp. 494-498.

Lambret-Frotte, J., Smith, G. and Langdale, J. A. (2024) 'GOLDEN2-like1 is sufficient but not necessary for chloroplast biogenesis in mesophyll cells of C4 grasses', *The Plant Journal,* 117(2), pp. 416-431.

Lamesch, P., Berardini, T. Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., Muller, R., Dreher, K., Alexander, D. L., Garcia-Hernandez, M., Karthikeyan, A. S., Lee, C. H., Nelson, W. D., Ploetz, L., Singh, S., Wensel, A. and Huala, E. (2012) 'The Arabidopsis Information Resource (TAIR): Improved gene annotation and new tools', *Nucleic Acids Research,* 40(Database issue).

Lin, M.-Y. (2020) *Studies into the genetic architecture of C 3 -C 4 characteristics in Moricandia* PhD Thesis, Heinrich-Heine-Universität Düsseldorf.

Liu, W. Y., Yu, C. P., Chang, C. K., Chen, H. J., Li, M. Y., Chen, Y. H., Shiu, S. H., Ku, M. S. B., Tu, S. L., Lu, M. Y. J. and Li, W. H. (2022) 'Regulators of early maize leaf development inferred from transcriptomes of laser capture microdissection (LCM)-isolated embryonic leaf cells', *Proceedings of the National Academy of Sciences of the United States of America,* 119(35), pp. e2208795119-e2208795119.

Lundgren, M. R. (2020) 'C2 photosynthesis: a promising route towards crop improvement?', *New Phytologist,* 228(6), pp. 1734-1740.

Lundgren, M. R. and Christin, P. A. (2017) 'Despite phylogenetic effects, C3-C4 lineages bridge the ecological gap to C4 photosynthesis', *Journal of experimental botany,* 68(2), pp. 241-254.

Mallmann, J., Heckmann, D., Bräutigam, A., Lercher, M. J., Weber, A. P. M., Westhoff, P. and Gowik, U. (2014) 'The role of photorespiration during the evolution of C4 photosynthesis in the genus Flaveria', *eLife,* 2014(3), pp. 1-23.

Monson, R. K. and Rawsthorne, S. (2000) 'CO2 Assimilation in C3-C4 Intermediate Plants', pp. 533-550.

Moreno-Villena, J. J., Zhou, H., Gilman, I. S., Lori Tausta, S., Maurice Cheung, C. Y. and Edwards, E. J. (2022) 'Spatial resolution of an integrated C4+CAM photosynthetic metabolism', *Science Advances,* 8(31), pp. 2349-2349.

Morgan, C. L., Turner, S. R. and Rawsthorne, S. (1993) 'Coordination of the cell-specific distribution of the four subunits of glycine decarboxylase and of serine hydroxymethyltransferase in leaves of C3-C4 intermediate species from different genera', *Planta,* 190(4), pp. 468-473.

Nobori, T., Monell, A., Lee, T. A., Zhou, J., Nery, J. and Ecker, J. R. (2023) 'Time-resolved single-cell and spatial gene regulatory atlas of plants under pathogen attack', *bioRxiv*, pp. 2023.04.10.536170-2023.04.10.536170.

Procko, C., Lee, T., Borsuk, A., Bargmann, B. O. R., Dabi, T., Nery, J. R., Estelle, M., Baird, L., O'Connor, C., Brodersen, C., Ecker, J. R. and Chory, J. (2022) 'Leaf cell-specific and single-cell transcriptional profiling reveals a role for the palisade layer in UV light protection', *The Plant Cell,* 34(9), pp. 3261-3279.

Rawsthorne, S. (1992) C3–C4 intermediate photosynthesis: linking physiology to gene expression. *The Plant Journal.*

Rawsthorne, S., Hylton, C. M., Smith, A. M. and Woolhouse, H. W. (1988a) 'Distribution of photorespiratory enzymes between bundle-sheath and mesophyll cells in leaves of the C3-C4 intermediate species Moricandia arvensis (L.) DC', *Planta,* 176(4), pp. 527-532.

Rawsthorne, S., Hylton, C. M., Smith, A. M. and Woolhouse, H. W. (1988b) 'Photorespiratory metabolism and immunogold localization of photorespiratory enzymes in leaves of C3 and C3-C4 intermediate species of Moricandia', *Planta,* 173(3), pp. 298-308.

Reeves, G., Grangé-Guermente, M. J. and Hibberd, J. M. (2017) 'Regulatory gateways for cell-specific gene expression in C4 leaves with Kranz anatomy', *Journal of Experimental Botany,* 68(2), pp. 107-116.

Sage, R. F. (2004) 'The evolution of C4 photosynthesis', *New Phytologist,* 161(2), pp. 341-370.

Sage, R. F. (2016) 'A portrait of the C4 photosynthetic family on the 50th anniversary of its discovery: Species number, evolutionary lineages, and Hall of Fame', *Journal of Experimental Botany,* 67(14), pp. 4039-4056.

Sage, R. F., Khoshravesh, R. and Sage, T. L. (2014) From proto-Kranz to C4 Kranz: Building the bridge to C 4 photosynthesis. *Journal of Experimental Botany.*

Sage, R. F., Monson, R. K., Ehleringer, J. R., Adachi, S. and Pearcy, R. W. (2018) 'Some like it hot: the physiological ecology of C4 plant evolution', *Oecologia 2018 187:4,* 187(4), pp. 941-966.

Sage, R. F., Sage, T. L. and Kocacinar, F. (2012) 'Photorespiration and the evolution of C4 photosynthesis', *Annual Review of Plant Biology,* 63, pp. 19-47.

Schlüter, U., Bräutigam, A., Gowik, U., Melzer, M., Christin, P. A., Kurz, S., Mettler-Altmann, T. and Weber, A. P. M. (2017) 'Photosynthesis in C3-C4 intermediate Moricandia species', *Journal of Experimental Botany,* 68(2), pp. 191-206.

Schlüter, U. and Weber, A. P. M. (2016) 'The Road to C4 Photosynthesis: Evolution of a Complex Trait via Intermediary States', *Plant and Cell Physiology,* 57(5), pp. 881-889.

Schulze, S., Westhoff, P. and Gowik, U. (2016) 'Glycine decarboxylase in C3, C4 and C3–C4 intermediate species', *Current Opinion in Plant Biology,* 31, pp. 29-35.

Selinski, J. and Scheibe, R. (2019) 'Malate valves: old shuttles with new perspectives', *Plant Biology (Stuttgart, Germany),* 21(Suppl Suppl 1), pp. 21-21.

Singh, P., Stevenson, S. R., Dickinson, P. J., Reyna-llorens, I., Tripathi, A., Reeves, G., Schreier, T. B. and Hibberd, J. M. (2023) 'C 4 gene induction during de-etiolation evolved through changes in cis to allow integration with ancestral C 3 gene regulatory networks', 9(13).

Slewinski, T. L., Anderson, A. A., Zhang, C. and Turgeon, R. (2012) 'Scarecrow Plays a Role in Establishing Kranz Anatomy in Maize Leaves', *Plant and Cell Physiology,* 53(12), pp. 2030-2037.

Steinegger, M. and Söding, J. (2017) 'MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets', *Nature Biotechnology 2017 35:11,* 35(11), pp. 1026-1028.

Stiehler, F., Steinborn, M., Scholz, S., Dey, D., Weber, A. P. M. and Denton, A. K. (2020) 'Helixer: Cross-species gene annotation of large eukaryotic genomes using deep learning', *Bioinformatics*.

Swift, J., Luginbuehl, L. H., Schreier, T. B., Donald, R. M., Lee, T. A., Nery, J. R., Ecker, J. R. and Hibberd, J. M. (2023) 'Single nuclei sequencing reveals C4 photosynthesis is based on rewiring of ancestral cell identity networks', *bioRxiv*, pp. 2023.10.26.562893-2023.10.26.562893.

Triesch, S., Denton, A. K., Bouvier, J. W., Buchmann, J. P., Reichel-Deland, V., Guerreiro, R. N. F. M., Busch, N., Schlüter, U., Stich, B., Kelly, S. and Weber, A. P. M. (2024) 'Transposable elements contribute to the establishment of the glycine shuttle in Brassicaceae species', *Plant Biology*, 26(2), pp. 270-281.

Walden, N., German, D. A., Wolf, E. M., Kiefer, M., Rigault, P., Huang, X. C., Kiefer, C., Schmickl, R., Franzke, A., Neuffer, B., Mummenhoff, K. and Koch, M. A. (2020) 'Nested whole-genome duplications coincide with diversification and high morphological disparity in Brassicaceae', *Nature Communications*, 11(1).

Walsh, C. A., Bräutigam, A., Roberts, M. R. and Lundgren, M. R. (2023) 'Evolutionary implications of C2 photosynthesis: How complex biochemical trade-offs may limit C4 evolution', *Journal of Experimental Botany*, 74(3), pp. 707-722.

Weber, A. P. M. and von Caemmerer, S. (2010) 'Plastid transport and metabolism of C3 and C4 plants — comparative analysis and possible biotechnological exploitation', *Current Opinion in Plant Biology*, 13(3), pp. 256-264.

Weckopp, S. C. and Kopriva, S. (2015) 'Are changes in sulfate assimilation pathway needed for evolution of c4 photosynthesis?', *Frontiers in Plant Science*, 5(JAN), pp. 124290-124290.

Williams, B. P., Aubry, S. and Hibberd, J. M. (2012) 'Molecular evolution of genes recruited into C 4 photosynthesis', *Trends in Plant Science*, 17(4), pp. 213-220.

Zhang, Y. L., Zhang, C. L., Wang, G. L., Wang, Y. X., Qi, C. H., Zhao, Q., You, C. X., Li, Y. Y. and Hao, Y. J. (2019) 'The R2R3 MYB transcription factor MdMYB30 modulates plant resistance against pathogens by regulating cuticular wax biosynthesis', *BMC plant biology*, 19(1).

“Some [DNA] may actually be genuine junk. And some (so the joke goes) may encode a message like *‘It was me, I'm God, I existed all along, ha ha’* ”

Terry Pratchett

8

Manuscript II

Transposable elements contribute to the establishment of the glycine shuttle in Brassicaceae species

Transposable elements contribute to the establishment of the glycine shuttle in Brassicaceae species

Sebastian Triesch[1,2], Alisandra K. Denton[1,2], Jacques W. Bouvier[5], Jan P. Buchmann[2,3], Vanessa Reichel-Deland[1], Ricardo Nuno Ferreira Martins Guerreiro[4], Noah Busch[1], Urte Schlüter[1,2], Benjamin Stich[2,4], Steven Kelly[5], Andreas P.M.Weber[1,2]

1 Institute for Plant Biochemistry, Heinrich Heine University Düsseldorf, Germany

2 Cluster of Excellence on Plant Sciences (CEPLAS)

3 Institute for Biological Data Sciences, Heinrich Heine University Düsseldorf, Germany

4 Institute for Quantitative Genetics and Genomics of Plants, Heinrich Heine University Düsseldorf, Germany

5 Department of Biology, University of Oxford, Oxford, United Kingdom

Author contributions

S.T. designed, performed and integrated all analyses. **J.W.B** and **S.K.** performed the phylogenetic correction of p-values. **N.B.** performed synteny analysis using *CoGe SynMap* under supervision of **S.T**.. **A.K.D.** performed gene annotations using *Helixer*. **A.P.M.W.**, **B.S.** and **U.S.** designed and coordinated the project. **A.K.D.**, **R.N.F.M.G.** and **B.S.** advised on statistical testing. All authors contributed to writing and accepted the manuscript.

Data availability

All data can be found in an Annotated Research Context Format (ARC) under https://git.nfdi4plants.org/ceplas/triesch2023_brassicaceae_transposons (public repository).

Publication

This work was originally published as:

Triesch *et al.,* "Transposable elements contribute to the establishment of the glycine shuttle in Brassicaceae species." *Plant Biology* 26.2 (2024): 270-281. https://doi.org/10.1111/plb.13601

The journal version of this paper can be found in chapter 11.

Abstract

C_3-C_4 intermediate photosynthesis has evolved at least five times convergently in the Brassicaceae, despite this family lacking *bona fide* C_4 species. The establishment of this carbon concentrating mechanism is known to require a complex suite of ultrastructural modifications as well as changes in spatial expression patterns, which are both thought to be underpinned by a reconfiguration of existing gene-regulatory networks. However, to date, the mechanisms which underpin the reconfiguration of these gene networks are largely unknown. In this study, we used a pan-genomic association approach to identify genomic features that could confer differential gene expression toward the C_3-C_4 intermediate state by analyzing eight C_3 species and seven C_3-C_4 species from five independent origins in the Brassicaceae. We found a strong correlation between transposable element (TE) insertions in *cis*-regulatory regions and C_3-C_4 intermediacy. Specifically, our study revealed 113 gene models in which the presence of a TE within a gene correlates with C_3-C_4 intermediate photosynthesis. In this set, genes involved in the photorespiratory glycine shuttle are enriched, including the glycine decarboxylase P-protein whose expression domain undergoes a spatial shift during the transition to C_3-C_4 photosynthesis. When further interrogating this gene, we discovered independent TE insertions in its upstream region which we conclude to be responsible for causing the spatial shift in *GLDP1* gene expression. Our findings hint at a pivotal role of TEs in the evolution of C_3-C_4 intermediacy, especially in mediating differential spatial gene expression.

Introduction

C_4 photosynthesis has convergently evolved more than 60 times in flowering land plants (Sage *et al.,* 2012). C_4 photosynthesis functions as a biochemical carbon concentrating mechanism that reduces the rate of photorespiration and thereby increases photosynthetic efficiency. Species that perform C_4 photosynthesis are mainly found in warm, dry and high-light environments in which leaf internal CO_2 are frequently low and by extension, the oxygenation to carboxylation ratio of Rubisco is elevated (Betti *et al.,* 2016; Sage *et al.,* 2012). Although C_4 photosynthesis has evolved independently in multiple disparate plant lineages, the complexity of the required anatomical, biochemical, and developmental adaptations makes engineering C_4 photosynthesis a difficult undertaking.

Plants that exhibit C_3-C_4 intermediate phenotypes are promising research subjects to study the early steps towards C_4 photosynthesis (Bellasio & Farquhar 2019; Kennedy & Laetsch 1974; Lundgren, 2020; Schlüter & Weber, 2016). C_3-C_4 intermediate species exhibit specialized anatomical traits and they differ from C_4 species as they do not possess a fully integrated C_4 cycle. C_3-C_4 intermediate traits are characterized by a lowered CO_2 compensation point (CCP), chloroplast and mitochondria-rich bundle-sheath cells (BSC) and, in some cases, an increased vein density (Christin *et al.,* 2011; Dengler *et al.,* 1994; Schlüter *et al.,* 2017). A further trait that is commonly shared between C_3-C_4 intermediate species from independent origins is the photorespiratory glycine shuttle, sometimes referred to as C_2 photosynthesis (reviewed in Schlüter & Weber, 2016). This shuttle relies on the BSC-specific decarboxylation of photorespiratory glycine, leading to an elevated CO_2 concentration around Rubisco. By extension, this increased partial pressure of CO_2 around the site of its fixation leads to a higher frequency of the Rubisco carboxylation reaction compared to oxygenation reactions, thereby suppressing photorespiration and resulting in decreased CCP (Kennedy & Laetsch, 1974; Monson & Edwards, 1984; Schlüter *et al.,* 2017).

Changes in the spatial and temporal patterns of gene expression are crucial for the evolution of C_3-C_4 intermediate photosynthesis (Hibberd & Covshoff 2010; Reeves *et al.,* 2017). Previously, it has been shown that the BSC-specific decarboxylation of glycine is caused by the differential localization of the glycine decarboxylase complex (GDC). In C_3-C_4 intermediate species from the genera *Moricandia*, *Flaveria* and *Panicum*, the P-protein of the GDC is only observed in BSC mitochondria, but not in mesophyll cell (MC) mitochondria (reviewed in Schulze *et al.,* 2016). This is a notable example of convergent evolution, as these species belong to the distant families Brassicaceae, Asteraceae and Poaceae, respectively. In these plants, loss of the GDC P-protein from the MC restricts glycine decarboxylation to the BSC in C_3-C_4 intermediate species (Morgan *et al.,* 1993; Rawsthorne *et al.,* 1988; Schulze *et al.,* 2016). However, the exact mechanism by which this is achieved differs between different species. For instance, in C_3 *Flaveria*, the gene encoding the GDC P-protein (*GLDP*) is present in two differentially regulated copies, *GLDPA* and *GLDPB*. In C_3-C_4 intermediate *Flaveria* species, the ubiquitously expressed *GLDPB* is downregulated compared to C_3 *Flaveria* species, whereas the BSC-specific *GLDPA* is highly expressed (Schulze *et al.,* 2013). In contrast, in C_3-C_4 intermediate *Moricandia,* the differential expression of *GLDP* is thought to be mediated by the loss of one gene copy and a change in regulation of the other copy. Specifically, in C_3-C_4 intermediate Brassiceae species, *GLDP2* is absent and *GLDP1* was reported to be differentially expressed by loss of a potential *cis*-element called M-Box. The M-Box element in the *Arabidopsis thaliana GLDP1* promoter confers a low-level expression in both MC and BSC and is absent from the upstream region of *GLDP1* in C_3-C_4 intermediate *Moricandia* species. A second *cis*-element, the V-Box, was shown to confer high levels of expression in the BSC and is present in all analyzed Brassicaceae *GLDP1* promoter sequences to date (Adwy *et al.,* 2015, 2019). Thus, there are multiple mechanisms through which *GLDP1* expression can be changed from being ubiquitously expressed in the leaf, to being BSC-specific in C_3-C_4 plants.

Structural variation can originate from the activity of mobile genetic elements. In plants, transposable elements (TEs) comprise a large fraction of mobile genetic elements and contribute substantially to genome size variation (Lee & Kim, 2014) and have substantial effects on the expression of genes (Hirsch & Springer, 2017). TEs can be divided into two classes (Wicker *et al.,* 2007) based on their transposition mechanisms: Class I transposons proliferate via a "copy-and-paste" mechanism involving an RNA intermediate, whereas Class II transposons transpose directly via a "cut-and-paste" mechanism. Due to their impact on structural variation, it has been frequently proposed that TEs can play a part in genome evolution and the evolution of novel genetic and phenotypic features (Buchmann *et al.,* 2012; Feschotte 2008; Qiu & Köhler, 2020; Wicker *et al.,* 2007). Decades ago, Britten & Davidson (1971) put forward the idea that the co-option of mobile sequences containing gene regulatory elements can connect genes to the same gene regulatory networks. The co-option of TEs for regulatory purposes is called "exaptation" (Brosius & Gould, 1992). In the present day with the vast amount of genomic data available, a deeper understanding of the role of transposable elements in genetic regulation allows linking genomic mechanisms with the evolution of complex traits.

TEs can rewire gene regulatory networks using different modes of action and influence the interplay of regulatory proteins (*trans*-elements) and the DNA sequences they are binding to (*cis*-elements). One such mode of action is the exaptation of a *cis*-regulatory element (CRE) from a separate gene (Fig 1). If the CRE inside a TE is copied from one gene and retained by the other gene, both genes become controlled by a mutual CRE and are thus connected by a shared gene regulatory network (Fig 1, B). In contrast to this scenario, it is also possible that TE integration into a CRE can suppress its function, either by interrupting the CRE sequence or altering the chromatin state of the respective CRE locus (Fig. 1, C) (Feschotte, 2008). A further possibility is the *de novo* generation of new CRE by point mutations in TEs (Fig. 1, D). New CREs, e.g. a 10-mer promoter element, can arise by random point mutations in between 700,000-4.8 million years (Behrens & Vingron, 2010).

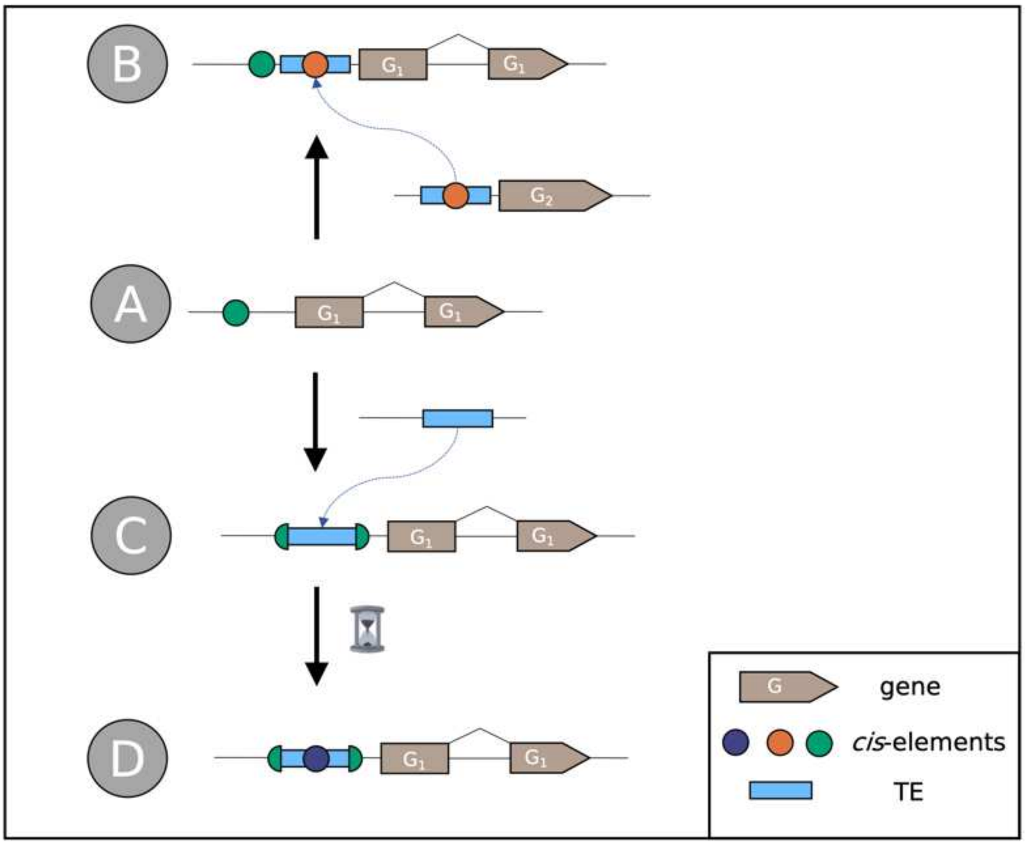


Figure 1: Schematic illustration of gene regulation rewiring by TE exaptation. **A**: The hypothetical gene G1 is controlled by a *cis*-regulatory element (CRE, green dot). **B**: Gene G2 is regulated by a different CRE (orange dot) located within a TE (blue box). Upon transposition of the TE to the upstream region of G1, G1 might co-opt the function of the orange CRE, thus connecting G1 and G2 to the same gene regulatory network. **C**: TE transposition can also lead to destruction or suppression of the CRE. **D**: During TE decay, new CREs (blue dot) might occur through accumulation of point mutations.

Several examples for the role of TEs in rewiring gene regulatory networks in plants have been reported. In rice, the *mPing* DNA transposon was found preferentially in the 5’ region and was associated with the upregulation of stress response genes (Naito *et al.,* 2009). In Brassicaceae, the evolution of heat-tolerance was linked to the activity of *Copia* retrotransposons containing heat-shock factor binding elements (Pietzenuk *et al.,* 2016). Furthermore, TEs were also found to be associated with endosperm development, e.g. the distribution of the PHEREs1 MADS-box transcription factor binding motifs by *Helitron* transposons in *A. thaliana* (Batista *et al.,* 2019). The *Youren* miniature inverted-repeat TE (*MITE*) was shown to be transcribed in rice endosperm, putatively mediated by a NUCLEAR FACTOR Y binding motif in the vicinity of the 5’ terminal inverted repeat (TIR) of *Youren* (Nagata *et al.,* 2022).

Previously, it has been shown that TEs play a significant role in the evolution of C_4 photosynthesis in maize. For instance, by analyzing 40 C_4 gene orthologs between rice and maize for the presence of BSC-specific promoter motifs, Cao *et al.* (2016) identified over 1,000 promoter motifs that were differentially distributed between C_3 and C_4 orthologs, of which more than 60 % were found to be associated with TEs and potentially co-opted by TE integration. These motifs may originate from non-photosynthetic genes and transposed to C_4 genes, which connected gene regulatory networks. The authors showed that TEs play a significant role in the evolution of C_4 photosynthesis in maize. However, the study of Cao *et al.* (2016) focused on evolutionary distant grasses, which makes it difficult to draw conclusions about the early evolutionary events towards C_4 photosynthesis.

In the present study, we test whether TE insertions are involved in decisive steps of the evolutionary establishment of C_3-C_4 intermediate photosynthesis. To do this, we focused on the Brassicaceae family which exhibits at least five independent origins of C_3-C_4 intermediate photosynthesis (Guerreiro *et al.,* 2023; Schlüter *et al.,* 2022) and contains multiple important and well-studied model plant species such as *A. thaliana*, *Arabis alpina* as well as relevant crop and vegetable plants such as *Brassica oleracea* and *Diplotaxis tenuifolia* (arugula).

We performed a pan-genomic association study to analyze the TE landscape of 15 Brassicaceae species. In doing so, we tested for correlations between TE positions and the presence of C_3-C_4 intermediate traits. Specifically, correlations between the presence or absence of upstream co-occurring TEs with the CO_2 compensation point were analyzed. In this unbiased approach, we aimed at finding genes that retained upstream TEs selectively only in C_3-C_4 intermediate plants. Based on the results of this analysis, we examined the upstream regions of relevant photorespiratory genes in closer detail to assess the potential role that TE insertions have played during establishment of C_3-C_4 photosynthesis traits. In doing so, we provide evidence that the insertion of TEs in *cis*-regulatory regions of key genes is associated with the evolution of C_3-C_4 photosynthesis in the Brassicaceae.

Material and Methods

Genomes and carbon compensation points

The genomes of *Brassica gravinae* (Bg), *Brassica tournefortii* (Bt), *Carrichtera annua* (Ca), *Diplotaxis erucoides* (De), *Diplotaxis tenuifolia* (Dt), *Diplotaxis viminea* (Dv), *Hirschfeldia incana* (accessions HIR1 and HIR3), *Moricandia nitens* (Mn) and *Moricandia suffruticosa* (Ms) were obtained from Guerreiro *et al.* (2023). *H. incana* HIR3 shows high phylogenetic similarity to *Sinapis pubescens* and the exact phylogenetic placement is unknown (Guerreiro *et al.,* 2023). For consistency, the name "*H. incana* HIR3" is retained throughout this work.

The genome of *Arabis alpina* (Aa) was obtained from Jiao *et al.* (2017). The genome of *Arabidopsis thaliana* (At) was obtained from Lamesch *et al.* (2012). The genome of *Moricandia arvensis* (Ma) and *Moricandia moricandioides* (Mm) were obtained from Lin *et al.* (2021). The genome assembly for *Brassica oleracea* (Bo) was obtained from Parkin *et al.* (2014). The genome for *Gynandropsis gynandra* (Gg) was obtained from Hoang *et al.* (2022). A full list of species names and accession number and sources can be found in Supplemental Table 1. Gas exchange data was obtained from Schlüter *et al.* (2022). The phylogenetic tree of all studied species was obtained from Guerreiro *et al.* (2023).

Gene annotation

Consistent structural gene annotations were generated for each species using *Helixer* (Holst *et al.,* 2023) with the hybrid convolutional and bidirectional long-short term memory model, HybridModel, specifically the trained instance of land_plant_v0.3_m_0100 with default parameters.

Annotation of transposable elements

TEs were *de novo* annotated using *EDTA* 1.9.9 (Ou *et al.,* 2019) using the -anno 1 and -sensitive 1 flags. For the calculation of genomic composition (Fig. 2, Fig. 3), intact and fragmented TEs were used. To reduce the influence of false-positive hits, the pan-genomic gene-TE association study was performed for intact TEs only. The long terminal repeats (LTR) insertion time was calculated using

$$t_{insertion} = \frac{1 - LTRidentity}{2 * \mu}$$

assuming a neutral mutation rate of $\mu = 1.4 \cdot 10^{-8}$ substitutions per site per year (Cai *et al.,* 2018). The LTR identity was calculated as fraction of conserved base pairs of the aligned LTRs from the identified LTR elements:

$$LTRidentity = \frac{number\ of\ conserved\ bp}{number\ of\ total\ bp}$$

Analysis of differential transposable element insertion

All downstream analyses were performed using *Python* 3.6 including *pandas* 1.2.4, *numpy* 1.20.1, *matplotlib* 3.4.1, *scikit-learn* 0.24.1, *scipy* 1.6.2 and *statsmodels* 0.12.2. The annotation files for genes and intact TEs were compared for each species. TEs were considered co-occurring with genes if their position matched one of the five cases described in Fig. 5. *CoGe SynMap* (https://genomevolution.org/coge/SynMap.pl) was used to identify orthologs and paralogs between the set of species. Each syntenic gene model was functionally annotated using *Mercator* 4.0 (Schwacke *et al.,* 2019).

For each obtained syntelog, the effect of the presence or absence of an upstream TE on CCP was assessed using a phylogenetic implementation of the one-way ANOVA which accounts for the non-independence between species on the phylogenetic tree. For this purpose, phylogenetic ANOVAs were performed in the R environment using the *phylANOVA* function in the *phytools* 1.0.3 package (Revell, 2012) using 1000 simulations and integrated posthoc comparisons to evaluate differences between means.

Enrichment of *Mercator* bins for genes with correlating upstream TEs was calculated using Fisher’s exact test. The identities of TEs in the *GLDP1* promoter were validated using the *CENSOR* webtool (Kohany *et al.,* 2006).

Results

The TE landscape of C_3 and C_3-C_4 Brassicaceae species

To screen for genomic features of potential relevance to the evolution of the C_3-C_4 photosynthesis trait, we conducted a pan-genomic association study of eight C_3 Brassicaceae species, seven C_3-C_4 intermediate Brassicaceae species from five independent origins, and one C_4 Cleomaceae as an outgroup species for tree building. The five independent origins of C_3-C_4 intermediate photosynthesis can be found in the *M. arvensis*, *M. nitens*, and *M. suffruticosa* monophylum, as well as in *D. erucoides*, *D. tenuifolia*, *B. gravinae*, and *H. incana* HIR3 (Fig. 2) (Guerreiro *et al.,* 2023; Schlüter & Weber 2016; Schlüter *et al.,* 2022).

The species panel exhibits genome sizes ranging from 120 Mbp in *A. thaliana* to 677 Mbp in *M. arvensis.* We found no significant difference in genome size between species exhibiting either the C_3 or C_3-C_4 intermediate photosynthesis phenotype (Fig 2; one-way ANOVA p>0.05). We next *de novo* annotated TEs using the *EDTA* pipeline (Ou *et al.,* 2019). Overall, the annotated fragmented and intact transposons made up between 18 % of the genome in *A. thaliana* and 75 % in *M. arvensis*. We observed differences in genome size and TE content also in closely related species, between *M. arvensis* and *M. moricandioides* and between *B. gravinae* and *D. viminea*. Furthermore, we observed that differences in genome size are mainly due to the different TE content.

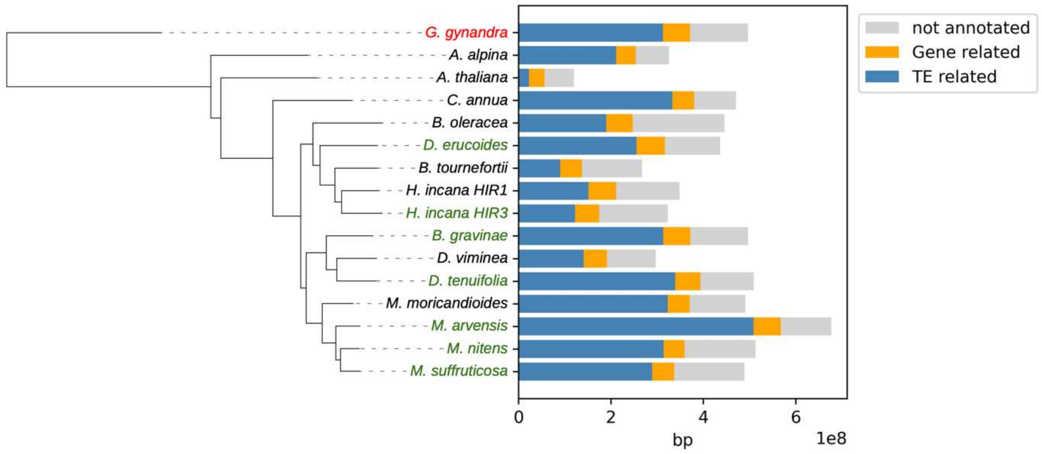


Figure 2: Phylogeny and genomic composition of 15 selected Brassicaceae species and the Cleomaceae outgroup. C_3-C_4 intermediate species are highlighted in green, the C_4 outgroup *G. gynandra* is highlighted in red. TE-related nucleotides are defined as spanning intact and fragmented transposons.

Class I type retrotransposons represented the majority of identified TEs across both C_3 and C_3-C_4 species (Fig. 3). For instance, across all analyzed genomes, between 60 % and 68 % of all annotated TEs were Class I retrotransposons. In contrast, the proportion of TE classes in the genomes varied greatly across species (Fig. 3; Supplemental Table S2). The TE Class II was dominated by TEs from the *Helitron* group, making up between five and 20 % of the genome (Fig. 3). The percentage of the genome made up of TEs from the different classes varied between the photosynthesis types, with a significantly higher amount of TEs in C_3-C_4 genomes (two-way ANOVA, p = 0.013).

To analyze recent increases of TE activity and their potential roles in the evolution of C_3-C_4 intermediate photosynthesis, we determined the insertion times of long terminal repeat (LTR) transposons (Fig. 4; Supplemental Table S3). *LTR retriever*, which is the LTR annotation tool of the *EDTA* pipeline, detected LTR transposons to a threshold for repeat identity of 91 %. Assuming a neutral mutation rate of $\mu = 1.4*10^{-8}$ substitutions per site per year (Cai *et al.,* 2018), LTR insertion times could thus be dated to a maximum of 4 million years ago. In general, both C_3 and C_3-C_4 intermediate species revealed the same broad pattern of LTR bursts. Specifically, in both groups, there was an increased frequency for LTR-TEs younger than two million years. However, the increase was more pronounced for C_3-C_4 intermediate species, largely on account of the high number of young LTR-TEs in *M. arvensis*. Statistical analysis revealed a significant correlation between the age distribution of LTR-transposons and the photosynthesis phenotype (two-way ANOVA, $p = 0.033$).

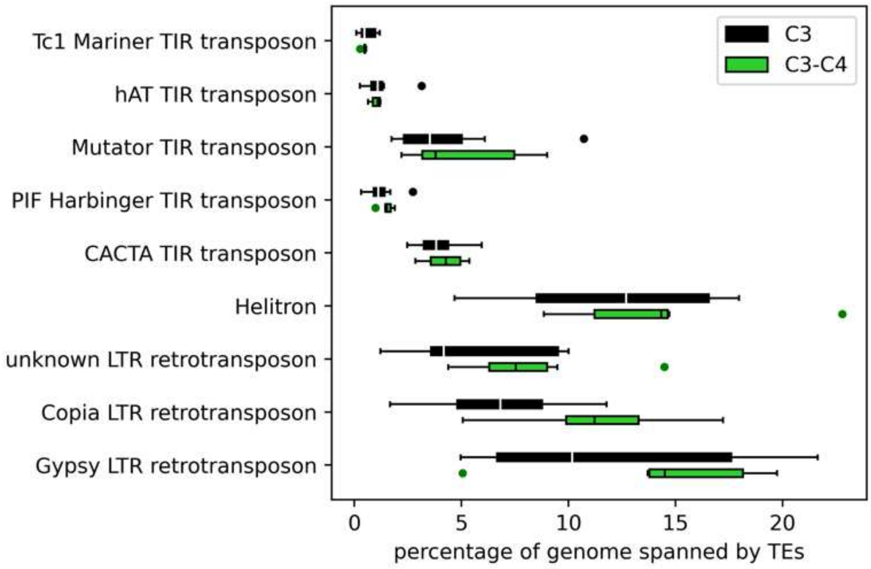


Figure 3: Boxplot indicating the percentage of the genome comprised by each class of intact and fragmented TEs in eight C_3 and six C_3-C_4 intermediate species. The y-axis shows the TE classes, the x-axis indicates the fraction of the genome made up by the respective TE class. Black boxes depict C_3 species and green boxes depict C_3-C_4 intermediate species.

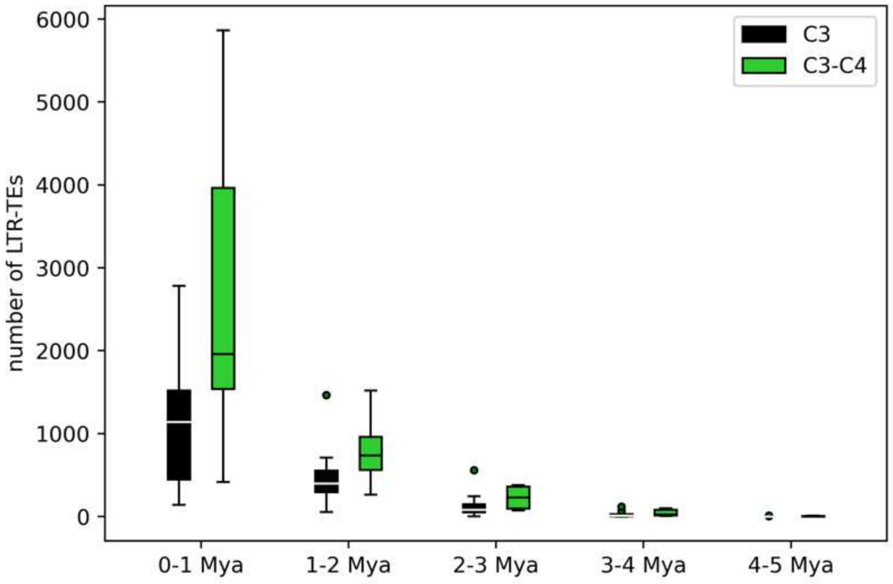


Figure 4: Boxplot of LTR-TE insertion times for eight C_3 and six C_3-C_4 intermediate species. The x-axis shows the insertion time in bins of 1 million years before today (Mya). The y-axis depicts the number of identified LTR-TEs calculated to be inserted within this timeframe. Calculation was performed using the LTR similarity of each LTR-TE and a neutral mutation rate of $1.4*10^{-8}$ substitutions per site per year. Black boxes represent C_3 species, green boxes represent C_3-C_4 species.

Upstream TEs are prevalent in C_3 and C_3-C_4 intermediate genomes

To better understand whether the high abundance of TEs in C_3-C_4 species was global or associated with specific genes, we next analyzed the differential co-occurrence of TEs with protein coding genes. Co-occurrent TEs were defined as follows (Fig. 5): (I) the TE starts or ends in a 3,000 bp window upstream of the gene (upstream), (II) the TE starts or ends in a 3,000 bp window downstream of the gene (downstream), (III) the TE is residing within an exon or intron of the gene (inside), (IV) the TE starts but only partially resides in the gene (start), or (V) the TE ends but only partially resides in the gene (end). Genes with TEs within the gene model (III) and overlapping TEs (IV and V) might have broken coding sequences and may result from imprecise annotations.

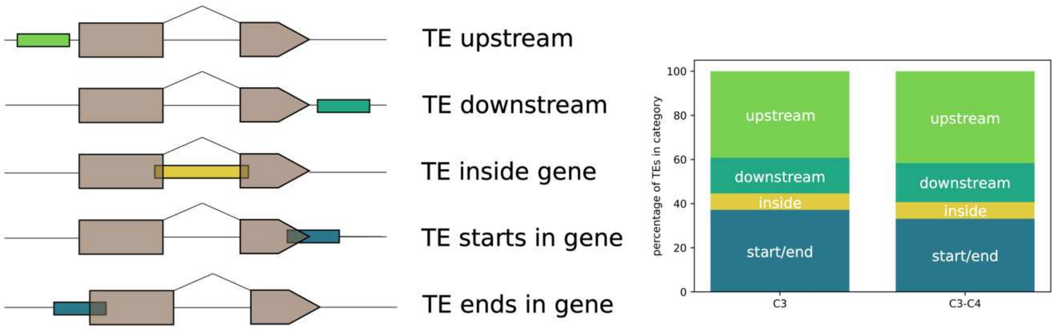


Figure 5: Left panel: Different contexts of TEs co-occurring with genes. **Right panel:** Bar charts indicating the fractions of TE co-occurring with genes within five contexts: starting or ending in a gene (start / end), residing within a gene (inside) or residing within a 3,000 bp window upstream or downstream the gene.

Table1: Selected subset of ten genes with upstream TEs with the lowest p-values for their association with the CCP.

Gene Name	AGI locus code	p-value
glycine dehydrogenase component P-protein of glycine cleavage system	AT4G33010	0.001
negative on TATA-less (NOT2)	AT5G59710	0.003
regulatory protein FLZ of SnRK1 complex	AT5G49120	0.004
pectate lyase	AT5G63180	0.005
MATE efflux family protein	AT2G38510	0.005
CYCLIN D-type regulatory protein	AT4G34160	0.005
regulatory protein FLZ of SnRK1 complex	AT5G47060	0.005
phosphocholine phosphatase (PS2/PECP1)	AT1G17710	0.007
PLATZ transcription factor family protein	AT3G50808	0.007
U-box domain-containing E3 ubiquitin ligase	AT4G25160	0.007

Analyzing potentially exaptated CREs, we focused on the up to 3,000 bp 5' region of the gene. To compare differential TE insertions between the analyzed species, we obtained syntenic gene information for *CoGe SynMap*. For each of these syntenic gene models, one-way ANOVA was employed, correlating the presence or absence of a co-occurring upstream TE with the CCP of the respective species. Across the selected eleven species, 55,148 TEs were identified to be co-occurring with a protein coding gene in at least one species, whereas 21,643 co-occurring TEs belonged to C$_3$ and 28,379 co-occurring TEs belonged to C$_3$-C$_4$ species. In both C$_3$ and C$_3$-C$_4$ intermediate species, over 50 % of the TEs co-occurring with genes were located up- or downstream of the gene (Fig. 5).

After correcting the p-values for the phylogenetic bias, we identified 113 genes where the co-occurrence of one of the gene with an upstream TE correlated with the CCP ($p \leq 0.05$; Tab. 1, Supplemental Table S5). Among the top ten genes (ranked by statistical confidence) were genes involved in photorespiration such as the genes encoding the T- and P-subprotein of the glycine decarboxylase complex (Fig. 6A). Strikingly, the C_3-C_4 intermediate orthologs of these genes exhibited upstream TEs, whereas the C_3 orthologs lacked upstream TEs. Thus, during the evolution of C_3-C_4, there was a "gain" in upstream TEs in genes that function in photorespiration (Fig. 6A). In the subset of genes which exhibit an association between the presence of an upstream TE and the plant CCP, two photorespiratory genes occurred (*GLDP*, *GLDT*). To quantify putative enrichment of certain gene ontologies, each gene was functionally annotated with a *Mercator* bin. Statistical enrichment analysis using Fisher's exact test revealed that the *Mercator* bin "Photosynthesis.Photorespiration" ($p = 0.0029$) was enriched in the set of genes that co-occur with upstream transposons (Tab. 2). The occurrence of this *Mercator* bin was increased 38-fold over the background, which is higher than for any other analyzed *Mercator* bin (Tab.2).

Table2: Results from two-sided Fisher's exact test for the enrichment of Mercator bins within the set of genes with significant upstream transposons.

***Mercator* Bin**	**p>0.05**	**p<0.05**	**p-value**	**Odds ratio**
Photosynthesis.Photorespiration	3	2	0.002907	38.2
Multi-process regulation.SnRK1-kinase regulation	9	2	0.014932	12.7
Cell wall organisation.cell wall proteins	32	3	0.022505	5.4
Solute transport.channels	45	3	0.050587	3.8

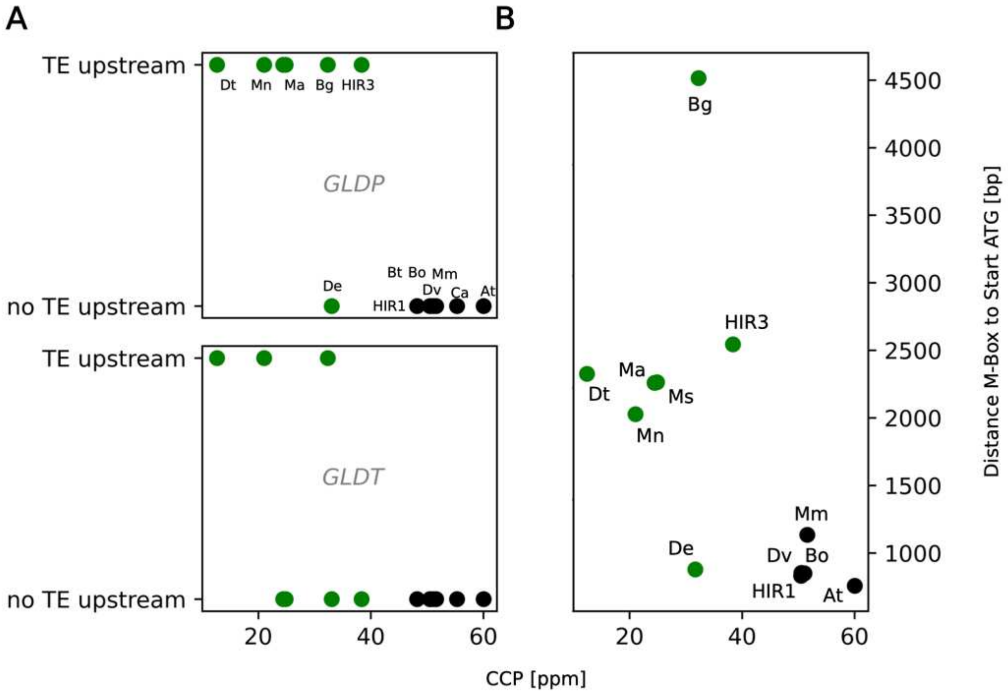


Figure 6: A: Scatter plot for two photorespiratory genes with significant co-associated upstream TEs. The y-axis indicates the presence of an upstream TE (yes/no), the x-axis shows the carbon compensation point. Abbreviations: *GLDP/GLDT*: P/T-protein of the GLYCINE DECARBOXYLASE COMPLEX **B: Scatter plot for the different architectures of the *GLDP1* promoter.** The y-axis indicates the distance between the conserved M-Box sequence and the *GLDP1* start site. Each dot represents a species. C_3 species are shown in green, C_3-C_4 intermediate species are shown in black. Species name abbreviations: At.: *Arabidopsis thaliana*, Bg: *Brassica gravinae*, Bo: *Brassica oleracea*, Bt: *Brassica tournefortii*, Ca: *Carrichtera annua*, De: *Diplotaxis erucoides*, Dt: *Diplotaxis tenuifolia*, Dv: *Diplotaxis viminea*, HIR1: *Hirschfeldia incana* HIR1, HIR3: *Hirschfeldia incana* HIR3, Ma: *Moricandia arvensis*, Mm: *Moricandia moricandioides*, Mn: *Moricandia nitens*, Ms: *Moricandia suffruticosa*.

The *GLDP1* upstream region shows independent TE insertions in C_3-C_4 intermediate genomes

As *GLDP* was the gene model with the strongest association between the presence of upstream TEs and CCP and it is known that the differential expression of *GLDP* contributes to the establishment of the photorespiratory glycine shuttle (Monson & Edwards, 1984; Rawsthorne *et al.,* 1988; Schulze *et al.,* 2013), we selected this gene for further analysis. Several studies about the underlying regulatory genetics of *GLDP* expression have been conducted before (Adwy *et al.,* 2015, 2019; Dickinson *et al.,* 2020; Schulze *et al.,* 2016).

Only one *GLDP* gene copy is present in species from the Brassiceae tribe that contains all known C_3-C_4 intermediate species of the Brassicaceae (Schlüter *et al.,* 2017). In contrast, the other two photorespiratory genes with correlating upstream TEs (Tab. 1; Fig. 6A) are found in higher copy numbers, which complicates a detailed genetic analysis.

We found three independent TE insertions in the promoter of C_3-C_4 intermediate *GLDP1* orthologs. In *D. tenuifolia* a *Mutator* TE starts at 1970 bp upstream of the *GLDP1* start codon. In *H. incana* HIR3 a TE of the *Helitron* class is located around 2240 bp upstream. In orthologs from the monophyletic clade *M. arvensis*, *M. nitens* and *M. suffruticosa* a *MITE* DNA transposon was detected, starting 1950 bp upstream of the *GLDP1* start codon. We calculated the minimum timespan since the *MITE* insertion by pairwise multiple sequence alignments of the *MITE* in the three *Moricandia GLDP1* promoters using the neutral mutation rate formula that was also employed for the calculation of LTR ages. We found that the *GLDP1* promoter *MITE* was at least 6.5 million years old.

All three independent TE insertions are located around 100 bp downstream of the M-Box promoter motif. This motif was previously hypothesized to confer MC expression (Adwy *et al.,* 2015) since truncation of the motif from the *AtGLDP1* promoter shifted GUS activity from the whole leaf blade to the veins. Furthermore, the M-Box was reported to be lost in C_3-C_4 intermediate *Moricandia* species (Adwy *et al.,* 2019). However, upon closer inspection, a highly conserved M-Box motif could be identified in all Brassicaceae genomes analyzed here. Notably, the M-Box was shifted upstream due to the TE insertion in C_3-C_4 species with the exception of *D. erucoides* (Fig. 6B, Fig. 7, Supplemental Table S6). In *B. gravinae*, the *EDTA* pipeline did not annotate an upstream transposon. However, we found a large insertion of unknown origin in the *B. gravinae GLDP1* promoter. This insertion is larger than the three reported TE cases but could be found in a similar position compared to the other *GLDP1* promoter insertions of TE origin (Fig. 7). In the *GLDP1* promoter of C_3-C_4 intermediate species *D. erucoides* no insertion could be found.

From five analyzed C_3-C_4 *GLDP1* promoters a large insertion behind the conserved M-Box could be found in four cases (monophyletic C_3-C_4 intermediate *Moricandia* clade, *D. tenuifolia*, *B. gravinae* and *H. incana* HIR3; Fig. 6B). Out of these four cases where the insertions occurred, we found evidence for the sequence being a TE in three cases (Fig. 7).

Discussion

Individual TE insertions, not global TE patterns are associated with C_3-C_4 intermediate photosynthesis

Evolution of new complex traits such as C_3-C_4 photosynthesis and C_4 photosynthesis requires the differential regulation of multiple genes. This includes differential gene expression across both MSC and BSC tissue as well as the installation of light-responsiveness for genes of the core metabolism (reviewed in Hibberd & Covshoff, 2010). In many cases, the evolution of differential gene regulation takes place in promoter sequences, either by introduction or suppression of *cis*-elements.

A few *cis*-elements for MC specificity have been previously found, including the MEM1 motif from the *Flaveria trinervia* phosphoenolpyruvate carboxylase gene (Gowik *et al.,* 2017) as well as the M-Box sequence in Brassicaceae (Adwy *et al.,* 2015; Dickinson *et* *al.,* 2020). TEs have the potential to deliver or suppress *cis*-elements upon their insertion in a target promoter. TEs can generate antisense transcription, interrupt or generate heterochromatic regions, or serve as raw material for the *de novo* evolution of new *cis*-elements (reviewed in Feschotte, 2008).

The role of TEs in the evolution of C_4 photosynthesis is only just started to being uncovered. The present study comprises the first pan-genomic association analysis to assess the importance of TEs in the evolution of C_3-C_4 intermediacy. Specifically, to do this, we analyzed the role of differential TE landscapes in 15 Brassicaceae species. Firstly, we investigated whether genome size and TE content correlate with the presence of the C_3-C_4 photosynthesis phenotype. Across our species panel a variety of genome sizes is present (Fig. 2), but we could detect no correlation between genome size and the presence of the photosynthesis trait. However, it is possible that different levels of heterozygosity in the sequenced species may confound these results and genome size estimations have to be handled with care.

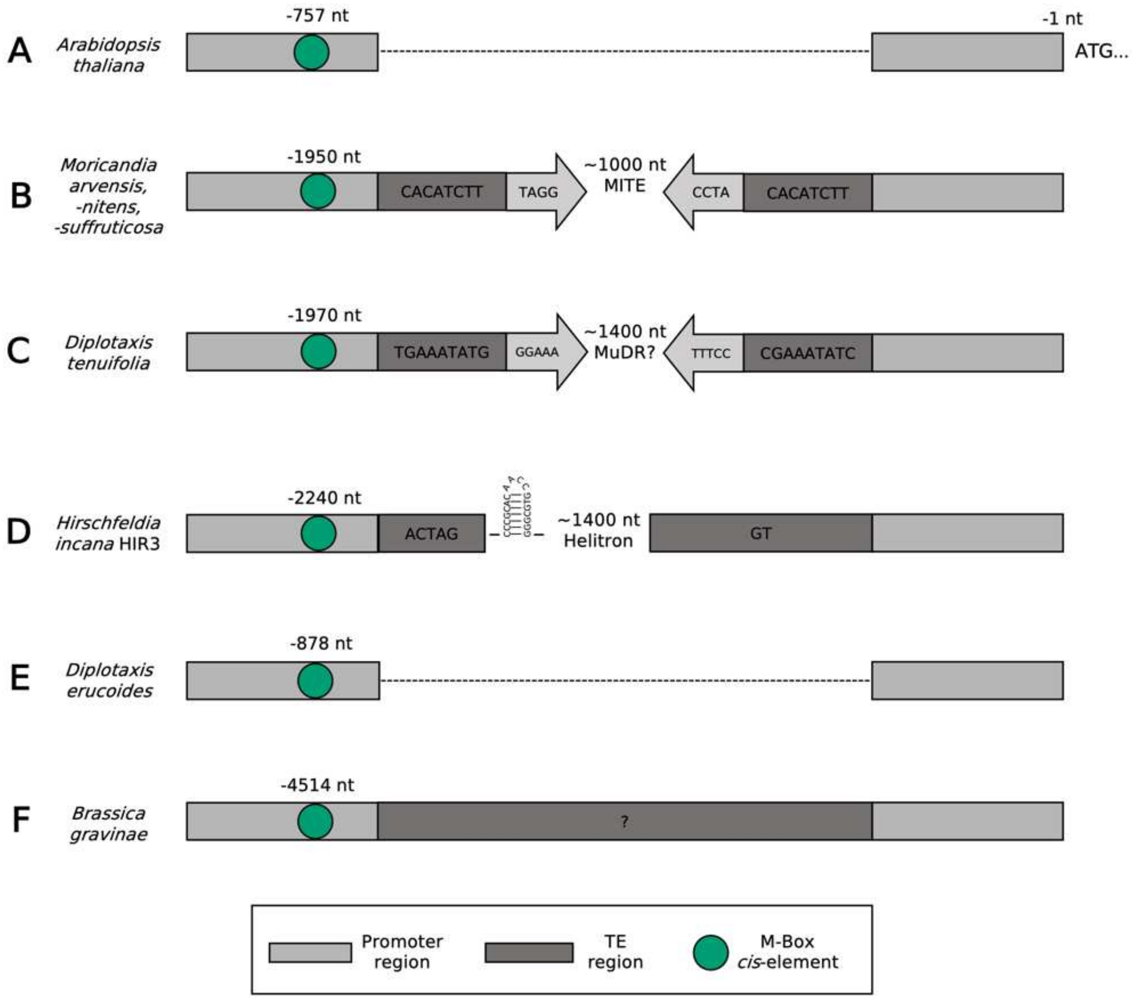


Figure 7: Schematic representation of the *GLDP1* promoter region. "ATG..." depicts the start site of the *GLDP1* gene. Dark grey boxes represent characteristic TE sites such as target site duplications or the *Helitron* insertion sites. Grey arrows depict terminal inverted repeat motifs. The M-Box motif is highlighted as a green circle. In C_3 species such as *Arabidopsis thaliana* no TE is annotated in the promoter sequence, leading to a low spacing between the M-Box and the *GLDP1* start site (**A**). In the C_3-C_4 intermediate *Moricandia* species, a *MITE* TE begins around 1,950 bp upstream of the *GLDP1* start codon (**B**). In *Diplotaxis tenuifolia*, a *Mutator* TE begins 1,970 bp upstream (**C**). In *Hirschfeldia incana* HIR3 a *Helitron* with a highly conserved hairpin loop structure is inserted around 2,240 bp upstream (**D**). Although being a C_3-C_4 intermediate species, the *D. erucoides GLDP1* promoter did not have an insertion behind the M-Box. (**E**). In *Brassica gravinae* a large insertion of unknown origin could be found behind the M-Box region (**F**).

Within the Brassicaceae family species exhibiting C_3-C_4 intermediate traits can only be found in the Brassiceae tribe. Notably, species from this tribe seem to have undergone recent polyploidization events (Walden *et al.,* 2020) and exhibit larger genome sizes than species from neighboring tribes (Lysak *et al.,* 2009).

Next, we analyzed the proportion of TEs across individual genomes. Our estimation of TE proportions is consistent with previously analyzed Brassicaceae genomes (Liu *et al.,* 2020; Mirouze & Vitte, 2014) and the *G. gynandra* genome (Hoang *et al.,* 2022). While genome size and TE content vary between species, we found a significant correlation between the photosynthesis phenotype and the proportion of the genome occupied by TEs in the respective species. Moreover, we found a recent burst in LTR-TE activity that is consistent with other studies (e.g. Cai *et al,* 2018). The recent sharp increase in LTR-TE bursts in C_3-C_4 species comes mainly from *M. arvensis* and might rather be due to high heterozygosity of LTR-containing genomic regions (Fig. 4). Although a significant correlation between LTR content and age with the C_3-C_4 intermediate phenotype could be identified, we cannot ultimately conclude that LTR transposon bursts contributed to the evolution of the C_3-C_4 intermediacy. Our LTR age analysis is limited to an LTR age of 4 million years. Given the estimated divergence time of 2-11 million years for C_3 and C_3-C_4 intermediate *Moricandia* species (Arias *et al.,* 2014), our analysis of LTR insertion times will miss the contribution of older LTRs to the evolution of C_3-C_4 intermediate traits. Furthermore, based on sequence identity between the C_3-C_4 intermediate *Moricandia GLDP1* promoters, we estimate the age of the *MITE* in the *Moricandia GLDP1* promoter to be at least 6.5 million years. This also falls within the proposed divergence time C_3 and C_3-C_4 intermediate *Moricandia* species of 2-11 million years (Arias, 2014). Thus, changes in TE content occurred concomitant with the evolution of C_3-C_4 intermediate photosynthesis and occurred in genes whose expression is required to change for operation of a C_3-C_4 cycle.

In the descriptive whole-genome view, we observed correlations between TE content and age and the C_3-C_4 intermediate phenotype. However, there is an individual TE pattern even in closely related lines (Fig. 2). We therefore conclude that the role of TE activity may have an influence on C_3-C_4 evolution, but not necessarily *via* means of general TE activity (TE outbursts or TE purging) but rather via selective TE insertions to relevant genes or upstream regions. To analyze this, we employed a pan-genomic *de novo* transposon-gene association study, where the co-occurrence of TEs with genes to the presence of a C_3-C_4 intermediate phenotype was correlated.

In both C_3 and C_3-C_4 intermediate species, more than 50 % of the analyzed co-occurring TEs were upstream or downstream of the respective co-occurring gene or spanning the gene. This is biologically plausible, as TEs crossing gene borders may disturb gene function and intergenic regions can harbor transposable elements (Buchmann *et al.,* 2012). Nevertheless, over 30 % of the transposons crossed the borders of annotated genes. We assume that this was due to imprecise annotations by the TE identification pipeline.

Differential gene regulation mediated by variation in upstream regions was shown to be a driver of C_4 trait evolution in multiple, well documented cases (Adwy *et al.,* 2015; Gowik *et al.,* 2017; Williams *et al.,* 2015; Wiludda *et al.,* 2012). Our analysis revealed 113 genes with an upstream TE that correlates with the presence of a C_3-C_4 intermediate phenotype (Fig. 7; $p<0.05$). Enrichment analysis of *Mercator* bins for this set of genes revealed an enrichment of the codes “Multi-process regulation.sucrose non-fermenting-related kinase (SnRK1) regulation” and “Photosynthesis.Photorespiration”. SnRK1 was shown to act as a central regulator of starvation metabolism that mediates energy homeostasis between organelles (Wurzinger *et al.,* 2018). During nutrient starvation, SnRK1 subcomplexes were found to regulate the differential expression of over 600 target genes (Baena-González *et al.,* 2007). Strikingly, ultrastructural adjustments and re-localization of the GDC P-protein to the BSC were demonstrated as a result of nitrogen starvation in the C_3-C_4 intermediate species *Chenopodium album* (Oono *et al.,* 2022).

There is a clear bias of TE retention upstream of photorespiratory and SnRK1-regulatory genes in C_3-C_4 intermediate species, although with a small effect size (2 out of 5 genes with $p<0.05$ for "Photosynthesis. Photorespiration"; 2 out of 11 genes with $p<0.05$ for "Multi-process regulation.SnRK1 regulation"; see Table 2).

We suggest that TE retention upstream of these genes has functional consequences such as differential gene expression, putatively due to the co-option of new, or, suppression of existing *cis*-elements. Strikingly, the set of genes that are significantly enriched for the presence of TEs in the upstream region contains multiple genes involved in photorespiration, such as those encoding the T- and P- proteins of the glycine decarboxylase complex (GLDT/GLDP). The modification of photorespiration is an important step towards the establishment of the glycine shuttle. The enrichment of TE insertions upstream of photorespiratory genes in C_3-C_4 intermediates is a potential hint that TEs play a significant role in the introduction of the glycine shuttle.

TE insertions in the *GLDP1* upstream region are highly convergent drivers of bundle-sheath cell specificity

GLDP is a well-characterized example for differential gene expression at the early stages of C_3-C_4 evolution across multiple plant lineages (Schlüter & Weber 2016; Schulze *et al.,* 2013). In the Brassiceae tribe, the *GLDP2* copy was lost (Schlüter *et al.,* 2017). Additionally, *GLDP1* was reported to be differentially expressed between C_3 and C_3-C_4 intermediate *Moricandia* species (Hylton *et al.,* 1988). In *A. thaliana*, GUS activity was restricted to the BSC by truncating the *GLDP1* promoter in the position of the M-Box, a promoter element approx. 800 bp upstream of the *AtGLDP1* gene start site. It was hypothesized that the M-Box confers MC expression, whereas expression in BSC is controlled by a MYC-MYB transcription factor binding module (Dickinson et al, 2023). Promoter-*GUS* fusions showed that the *GLDP1* promoter of the C_3 species *M. moricandioides* conferred *GUS* expression to both MC and BSC, whereas the *GLDP1* promoter of the C_3-C_4 intermediate species *M. arvensis* restricted *GUS* expression to the BSC (Adwy *et al.,* 2019).

Adwy *et al.* (2019) explain the establishment of the glycine shuttle in *Moricandia* by the loss of the M-Box in C_3-C_4 intermediate *Moricandia* species. In contrast, we found the M-Box sequence in all our analyzed *GLDP1* promoter variants, though this motif was shifted by over 1,000 bp further upstream by the insertion of three independent TEs in the promoters in three independent evolutionary origins of C_3-C_4 intermediate photosynthesis, and by an insertion of unknown provenance in a fourth independent origin. This shift may have led to the M-Box being overlooked in previous studies.

Based on the findings by Adwy *et al.* (2019) we conclude that not the loss of the M-Box, but rather the upstream shift of the element by insertion of a TE has led to the differential tissue specific expression of the *GLDP1* gene. The upstream shift of the M-Box was mediated by three independent TE insertions in lines with independent evolutionary origins of C_3-C_4 photosynthesis. This hints at a remarkable convergent evolutionary genetic mechanism in C_3-C_4 evolution. We suggest that the loss of *GLDP2* paved the way for neofunctionalization of the *GLDP1* copy in the Brassiceae tribe, the only Brassicaceae tribe containing C_3-C_4 intermediate species. This was mediated by the insertion of a TE in the promoter, suppressing the M-Box element and shifting *GLDP1* expression. It is questionable whether the TE insertion took place before or after the preconditioning of C_3-C_4 photosynthesis by anatomical adaptations such as higher vein density and the distinct leaf anatomy. Hypothetically, limited expression of *GLDP1* in the MC may have been deleterious without further adaptations, which could have prevented the TE retention in the promoter. In *D. erucoides* we do not find a transposon in the *GLDP1* promoter region. The spacing of the M-Box to the *GLDP1* start codon is in the range of C_3 plants (Fig. 6B). However, *D. erucoides* shows C_3-C_4 intermediate phenotypes (Lundgren, 2020; Schlüter *et al.,* 2017). We assume that, being an independent evolutionary origin of C_3-C_4 intermediate photosynthesis, *D. erucoides* either shifted *GLDP1* expression to the BSC by different means or, alternatively, that there must be other additional regulators in the *GLDP1* promoter beyond our transposon-M-Box model. Contrasting the well-studied GDC activity and localization in *Moricandia* species, there is no data on the *D. erucoides* GDC biochemistry and genetics. Therefore, we cannot rule out that the glycine shuttle in *D. erucoides* is mediated by a different GDC regulation compared to the other C_3-C_4 intermediate species, such as the differential activity of the GDC T-, L- or H- proteins. By adopting a whole-genome view of TE density and gene-TE associations, our study highlights the potential importance of TE insertions in contributing to the convergent evolution of C_3-C_4 intermediacy. Differential *GLDP* expression is one of the most important innovations that occurs and facilitates the establishment of the glycine shuttle. The novel genetic mechanism of differential *GLDP1* regulation by a TE-mediated insertion causing an upstream shift of the M-Box must be verified in experimental work. The lack of efficient transformation protocols represents a significant impediment to functional genetics studies in non-model plants. Thus far, the successful transformation of any plant within our Brassicaceae species panel, apart from *A. thaliana*, has proven elusive, thereby precluding genomic engineering in C_3-C_4 intermediate Brassicaceae. The validation of the impact of TEs, for example on *GLDP1* expression *in planta*, hinges on the future accessibility of these species to genetic transformation. These experiments may necessitate the alteration of TE types or manipulating the positioning of CREs in upstream regions. For example, using a CRISPR-associated genomic engineering technique, TE insertions in upstream regions could be changed to different TE types, elongated, shortened or even relocated to downstream or intronic positions. Studying the influence of TEs on regulatory upstream regions via promoter-reporter studies can be conducted using transgenic *A. thaliana* lines. Nonetheless, it is imperative to consider that, due to their involvement in epigenetic regulation, particularly as hotspots for cytosine methylation, transgenic TEs may behave distinctly in transgenic *A. thaliana* when compared to their behavior in their native host plant. Studying those genetic mechanisms of gene regulation in C_3-C_4 intermediate species will pave the way for a better understanding of the C_4 trait and facilitate genetic engineering efforts.

Supplementary Material

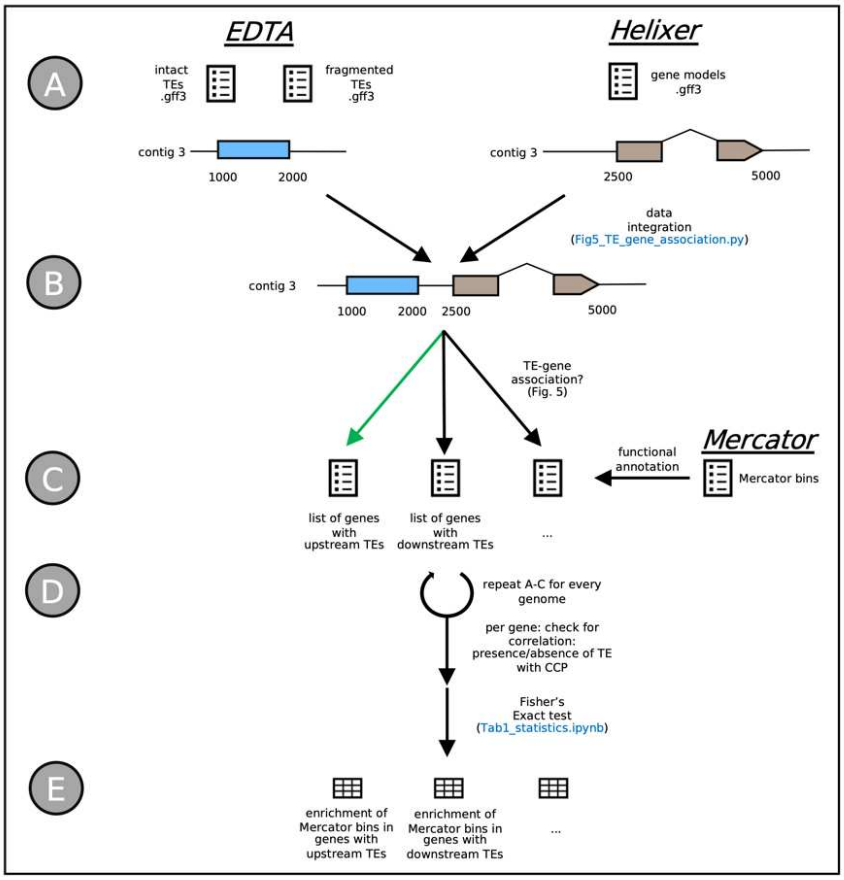
All supplemental tables can be found under:

https://git.nfdi4plants.org/hhu-plant-biochemistry/triesch2023_brassicaceae_transposons

Supplementary Table S1: Overview over selected species with photosynthesis type and accession number or source.

Supplementary Table S2: Number of nt spanned by intact and fragmented transposable elements per species analyzed.

Supplementary Table S3: Insertion times (age) of long terminal repeat transposons for each analyzed species.


EDTA
Helixer
A
intact TEs .gff3
fragmented TEs .gff3
gene models .gff3
contig 3
1000
2000
contig 3
2500
5000
data integration (Fig5_TE_gene_association.py)
B
contig 3
1000
2000
2500
5000
TE-gene association? (Fig. 5)
Mercator
functional annotation
Mercator bins
C
list of genes with upstream TEs
list of genes with downstream TEs
...
D
repeat A-C for every genome
per gene: check for correlation: presence/absence of TE with CCP
Fisher's Exact test (Tab1_statistics.ipynb)
E
enrichment of Mercator bins in genes with upstream TEs
enrichment of Mercator bins in genes with downstream TEs
...

Supplementary Figure S4: Flow chart depicting the computational workflow for the pan-genomic transposon-gene association study. File names highlighted in blue refer to scripts under https://git.nfdi4plants.org/hhu-plant-biochemistry/triesch2023_brassicaceae_transposons/-/tree/main/workflows. **A:** The *extensive de-novo TE annotator* (*EDTA*) software was used to annotate transposons in the selected genome sequences. EDTA distinguishes between intact and fragmented transposable elements (TEs). For the correlation of TEs and genes, only the intact TEs were used. Illustrated is one example TE (blue box) on a hypothetical contig at position 1,000-2,000 on the contig. *Helixer* was used to generate structural gene annotations. Depicted is one example gene (brown boxes) on a hypothetical contig at position 2,500-5,000 on the contig. **B:** Using a custom *python* script, the .gff3 files, containing the TE and gene annotations were compared and TE-gene associations as depicted in Fig. 5 were searched. In the example, the TE (blue box) resides up to 500 bp upstream of the example gene (brown box) and would thus be considered an upstream TE. **C:** For each genome, lists containing genes with TEs from the categories presented in Fig. 5 were created. The example from **B** would thus be appended to a list with genes that are associated with upstream TEs. *Mercator* was used to assign a functional annotation (*Mercator* bin) to all genes. Steps **A-C** were repeated for each genome. **D:** From the lists of genes with associated TEs per genome, a matrix was created where for each gene and species, the association of a gene with a TE was correlated with the carbon compensation point (CCP) of the species. These associations were tested using one-way ANOVA and resulting p-values were corrected for phylogenetic bias. Thus, a corrected p-value was assigned to each gene that indicated, whether there was a correlation of an associated TE with the CCP. **E:** From the p-values per gene, an arbitrary threshold of $p < 0.05$ was applied to divide the dataset. Fisher's test was used to quantify enrichment of *Mercator* bins within genes with $p < 0.05$.

Supplementary Table S4: Results of pan-genomic gene-transposon association study. Per gene, the absence (0) or presence (1) of a transposon within 3000 bp upstream of a gene is indicated for each analyzed species. The AGI code represents the *A. thaliana* gene with the highest sequence homology.

Supplementary Table S5: Distance of the M-Box to the *GLDP1* transcriptional start site for each analyzed *GLDP1* ortholog upstream region.

Literature

Adwy, W., Laxa, M. & Peterhansel, C. (2015), 'A simple mechanism for the establishment of C2-specific gene expression in Brassicaceae', *Plant Journal* **84**(6), 1231–1238.

Adwy, W., Schlüter, U., Papenbrock, J., Peterhansel, C. & Offermann, S. (2019), 'Loss of the M-box from the glycine decarboxylase P-subunit promoter in C2 Moricandia species', *Plant Gene* **18**.

Arias, T., Beilstein, M. A., Tang, M., McKain, M. R. & Pires, J. C. (2014), 'Diversification times among Brassica (Brassicaceae) crops suggest hybrid formation after 20 million years of divergence', *American journal of botany* **101**(1), 86–91.

Baena-González, E., Rolland, F., Thevelein, J. M. & Sheen, J. (2007), 'A central integrator of transcription networks in plant stress and energy signalling', *Nature 2007 448:7156* **448**(7156), 938–942.

Batista, R. A., Moreno-Romero, J., Qiu, Y., van Boven, J., Santos-González, J., Figueiredo, D. D. & Köhler, C. (2019), 'The mads-box transcription factor pheres1 controls imprinting in the endosperm by binding to domesticated transposons', *eLife* **8**.

Behrens, S. & Vingron, M. (2010), 'Studying the evolution of promoter sequences: A waiting time problem', *Journal of Computational Biology* **17**(12), 1591–1606.

Bellasio, C. & Farquhar, G. D. (2019), 'A leaf-level biochemical model simulating the introduction of C2 and C4 photosynthesis in C3 rice: gains, losses and metabolite fluxes', *New Phytologist* .

Betti, M., Bauwe, H., Busch, F. A., Fernie, A. R., Keech, O., Levey, M., Ort, D. R., Parry, M. A., Sage, R., Timm, S., Walker, B. & Weber, A. P. (2016), 'Manipulating photorespiration to increase plant productivity: Recent advances and perspectives for crop improvement', *Journal of Experimental Botany* **67**(10), 2977–2988.

Britten, R. J. & Davidson, E. H. (1971), 'Repetitive and non-repetitive DNA sequences and a speculation on the origins of evolutionary novelty.', *The Quarterly review of biology* **46**(2), 111–138.

Brosius, J. & Gould, S. J. (1992), 'On 'genomenclature': A comprehensive (and respectful) taxonomy for pseudogenes and other 'junk DNA'', *Proceedings of the National Academy of Sciences of the United States of America* **89**(22), 10706–10710.

Buchmann, J. P., Matsumoto, T., Stein, N., Keller, B. & Wicker, T. (2012), 'Inter-species sequence comparison of Brachypodium reveals how transposon activity corrodes genome colinearity', *The Plant Journal* **71**(4), 550–563.

Cai, X., Cui, Y., Zhang, L., Wu, J., Liang, J., Cheng, L., WANG, X. & Cheng, F. (2018), 'Hotspots of Independent and Multiple Rounds of LTR-retrotransposon Bursts in Brassica Species', *Horticultural Plant Journal* **4**(4), 165–174.

Cao, C., Xu, J., Zheng, G. & Zhu, X. G. (2016), 'Evidence for the role of transposons in the recruitment of cis-regulatory motifs during the evolution of C4 photosynthesis', *BMC Genomics* **17**(1).

Christin, P. A., Sage, T. L., Edwards, E. J., Ogburn, R. M., Khoshravesh, R. & Sage, R. F. (2011), 'Complex evolutionary transitions and the significance of c3–c4 intermediate forms of photosynthesis in molluginaceae', *Evolution* **65**(3), 643–660.

Dengler, N. G., Dengler, R. E., Donnelly, P. M. & Hattersley, P. W. (1994), 'Quantitative Leaf Anatomy of C3 and C4 Grasses (Poaceae): Bundle Sheath and Mesophyll Surface Area Relationships', *Annals of Botany* **73**(3), 241–255.

Dickinson, P. J., Knerovà, J., Szecowka, M., Stevenson, S. R., Burgess, S. J., Mulvey, H., Bagman, A. M., Gaudinier, A., Brady, S. M. & Hibberd, J. M. (2020), 'A bipartite transcription factor module controlling expression in the bundle sheath of Arabidopsis thaliana', *Nature Plants* **6**(12), 1468–1479.

Dickinson, P. J., Triesch, S., Schlüter, U., Weber, A. P., & Hibberd, J. M. (2023). A transcription factor module mediating C2 photosynthesis. *bioRxiv*, 2023-09.

Feschotte, C. (2008), 'Transposable elements and the evolution of regulatory networks', *Nature Reviews Genetics* **9**(5), 397–405.

Gowik, U., Schulze, S., Saladi´e, M., Rolland, V., Tanz, S. K., Westhoff, P. & Ludwig, M. (2017), 'A MEM1-like motif directs mesophyll cell-specific expression of the gene encoding the C4 carbonic anhydrase in Flaveria', *Journal of Experimental Botany* **68**(2), 311–320.

Guerreiro, R., Bonthala, V. S., Schlüter, U., Hoang, N. V., Triesch, S., Schranz, M. E. (2023) 'A genomic panel for studying C3-C4 intermediate photosynthesis in the Brassiceae tribe', *Plant, Cell & Environment*, 1–17. *https://doi.org/10.1111/pce.14662*

Hibberd, J. M. & Covshoff, S. (2010), 'The Regulation of Gene Expression Required for C4 Photosynthesis', *http://dx.doi.org/10.1146/annurev-arplant-042809-112238* **61**, 181–207.

Hirsch, C. D. & Springer, N. M. (2017), 'Transposable element influences on gene expression in plants', *Biochimica et biophysica acta. Gene regulatory mechanisms* **1860**(1), 157–165.

Hoang, N. V., Sogbohossou, E. O. D., Xiong, W., Simpson, C. J. C., Singh, P., van den Bergh, E., Zhu, X.-G., Brautigam, A., Weber, A. P. M., van Haarst, J. C., Schijlen, E. G. W. M., Hendre, P. S., Deynze, A. V., Achigan-Dako, E. G., Hibberd, J. M. & Schranz, M. E. (2022), 'The genome of Gynandropsis gynandra provides insights into whole-genome duplications and the evolution of C4 photosynthesis in Cleomaceae', *bioRxiv* p. 2022.07.09.499295.

Holst, F., Bolger, A., Günther, C., Maß, J., Triesch, S., Kindel, F., Kiel, N., Saadat, N., Ebenhöh, O., Usadel, B., Schwacke, R., Bolger, M., Weber, A. P. & Denton, A. K. (2023), 'Helixer–de novo Prediction of primary eukaryotic gene models conbining deep learning and a Hidden Marcov Model', *bioRxiv* .

Hylton, C. M., Rawsthorne, S., Smith, A. M., Jones, D. A. & Woolhouse, H. W. (1988), 'Glycine decarboxylase is confined to the bundle-sheath cells of leaves of C3-C4 intermediate species', *Planta* **175**(4), 452–459.

Jiao, W. B., Accinelli, G. G., Hartwig, B., Kiefer, C., Baker, D., Severing, E., Willing, E. M., Piednoel, M., Woetzel, S., Madrid-Herrero, E., Huettel, B., Hümann, U., Reinhard, R., Koch, M. A., Swan, D., Clavijo, B., Coupland, G. & Schneeberger, K. (2017), 'Improving and correcting the contiguity of long-read genome assemblies of three plant species using optical mapping and chromosome conformation capture data', *Genome Research* **27**(5), 778–786.

Kennedy, R. A. & Laetsch, W. M. (1974), 'Plant species intermediate for C3, C4 photosynthesis', *Science* **184**(4141), 1087–1089.

Kohany, O., Gentles, A. J., Hankus, L. & Jurka, J. (2006), 'Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor', *BMC Bioinformatics* **7**.

Lamesch, P., Berardini, T. Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., Muller, R., Dreher, K., Alexander, D. L., Garcia-Hernandez, M., Karthikeyan, A. S., Lee, C. H., Nelson, W. D., Ploetz, L., Singh, S., Wensel, A. & Huala, E. (2012), 'The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools', *Nucleic acids research* **40**(Database issue).

Lee, S.-I. & Kim, N.-S. (2014), 'Transposable Elements and Genome Size Variations in Plants', *Genomics & Informatics* **12**(3), 87.

Lin, M.-Y., Koppers, N., Denton, A., Schlüter, U. & Weber, A. P. (2021), 'Whole genome sequencing and assembly data of Moricandia moricandioides and M. arvensis', *Data in Brief* **35**, 106922.

Liu, Z., Fan, M., Yue, E. K., Li, Y., Tao, R. F., Xu, H. M., Duan, M. H. & Xu, J. H. (2020), 'Natural variation and evolutionary dynamics of transposable elements in Brassica oleracea based on next-generation sequencing data', *Horticulture Research* **7**(1).

Lundgren, M. R. (2020), 'C2 photosynthesis: a promising route towards crop improvement?', *New Phytologist* **228**(6), 1734–1740.

Lysak, M. A., Koch, M. A., Beaulieu, J. M., Meister, A. & Leitch, I. J. (2009), 'The dynamic ups and downs of genome size evolution in Brassicaceae', *Molecular Biology and Evolution* **26**(1), 85–98.

Mirouze, M. & Vitte, C. (2014), 'Transposable elements, a treasure trove to decipher epigenetic variation: insights from Arabidopsis and crop epigenomes', *Journal of Experimental Botany* **65**(10), 2801–2812.

Monson, R. K. & Edwards, G. E. (1984), ' C 3 - C 4 Intermediate Photosynthesis in Plants ', *BioScience* **34**(9), 563–574.

Morgan, C. L., Turner, S. R. & Rawsthorne, S. (1993), 'Coordination of the cell-specific distribution of the four subunits of glycine decarboxylase and of serine hydroxymethyltransferase in leaves of C3-C4 intermediate species from different genera', *Planta* **190**(4), 468–473.

Nagata, H., Ono, A., Tonosaki, K., Kawakatsu, T., Sato, Y., Yano, K., Kishima, Y. & Kinoshita, T. (2022), 'Temporal changes in transcripts of miniature inverted-repeat transposable elements during rice endosperm development', *Plant Journal* **109**(5), 1035–1047.

Naito, K., Zhang, F., Tsukiyama, T., Saito, H., Hancock, C. N., Richardson, A. O., Okumoto, Y., Tanisaka, T. & Wessler, S. R. (2009), 'Unexpected consequences of a sudden and massive transposon amplification on rice gene expression', *Nature* **461**(7267), 1130–1134.

Oono, J., Hatakeyama, Y., Yabiku, T. & Ueno, O. (2022), 'Effects of growth temperature and nitrogen nutrition on expression of C3–C4 intermediate traits in Chenopodium album', *Journal of Plant Research* .

Ou, S., Su, W., Liao, Y., Chougule, K., Agda, J. R., Hellinga, A. J., Lugo, C. S. B., Elliott, T. A., Ware, D., Peterson, T., Jiang, N., Hirsch, C. N. & Hufford, M. B. (2019), 'Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline', *Genome Biology* **20**(1).

Parkin, I. A., Koh, C., Tang, H., Robinson, S. J., Kagale, S., Clarke, W. E., Town, C. D., Nixon, J., Krishnakumar, V., Bidwell, S. L., Denoeud, F., Belcram, H., Links, M. G., Just, J., Clarke, C., Bender, T., Huebert, T., Mason, A. S., Chris Pires, J., Barker, G., Moore, J., Walley, P. G., Manoli, S., Batley, J., Edwards, D., Nelson, M. N., Wang, X., Paterson, A. H., King, G., Bancroft, I., Chalhoub, B. & Sharpe, A. G. (2014), 'Transcriptome and methylome profiling reveals relics of genome dominance in the mesopolyploid Brassica oleracea', *Genome biology* **15**(6).

Pietzenuk, B., Markus, C., Gaubert, H., Bagwan, N., Merotto, A., Bucher, E. & Pecinka, A. (2016), 'Recurrent evolution of heat-responsiveness in Brassicaceae COPIA elements', *Genome Biology* **17**(1)

Qiu, Y. & Köhler, C. (2020), 'Mobility connects: Transposable elements wire new transcriptional networks by transferring transcription factor binding motifs', *Biochemical Society Transactions* **48**(3), 1005–1017.

Rawsthorne, S., Hylton, C. M., Smith, A. M. & Woolhouse, H. W. (1988), 'Photorespiratory metabolism and immunogold localization of photorespiratory enzymes in leaves of C3 and C3-C4 intermediate species of Moricandia', *Planta* **173**(3), 298–308.

Reeves, G., Grangè-Guermente, M. J. & Hibberd, J. M. (2017), 'Regulatory gateways for cell-specific gene expression in C4 leaves with Kranz anatomy', *Journal of Experimental Botany* **68**(2), 107–116.

Revell, L. J. (2012), 'phytools: An R package for phylogenetic comparative biology (and other things)', *Methods in Ecology and Evolution* .

Sage, R. F., Sage, T. L. & Kocacinar, F. (2012), 'Photorespiration and the evolution of C4 photosynthesis', *Annual Review of Plant Biology* **63**, 19–47.

Schlüter, U., Bräutigam, A., Gowik, U., Melzer, M., Christin, P. A., Kurz, S., Mettler-Altmann, T. & Weber, A. P. (2017), 'Photosynthesis in C3-C4 intermediate Moricandia species', *Journal of Experimental Botany* **68**(2), 191–206.

Schlüter, U. & Weber, A. P. (2016), 'The Road to C4 Photosynthesis: Evolution of a Complex Trait via Intermediary States', *Plant and Cell Physiology* **57**(5), 881–889.

Schlüter, U., Bouvier, J. W., Guerreiro, R., Malisic, M., Kontny, C., Westhoff, P., Stich, B. & Weber, A. P. M. (2022), 'Brassicaceae display diverse photorespiratory carbon recapturing mechanisms', *bioRxiv*

Schulze, S., Mallmann, J., Burscheidt, J., Koczor, M., Streubel, M., Bauwe, H., Gowik, U. & Westhoff, P. (2013), 'Evolution of C4 photosynthesis in the genus flaveria: Establishment of a photorespiratory CO2 pump', *Plant Cell* **25**(7), 2522–2535.

Schulze, S., Westhoff, P. & Gowik, U. (2016), 'Glycine decarboxylase in C3, C4 and C3–C4 intermediate species', *Current Opinion in Plant Biology* **31**, 29–35.

Schwacke, R., Ponce-Soto, G. Y., Krause, K., Bolger, A. M., Arsova, B., Hallab, A., Gruden, K., Stitt, M., Bolger, M. E. & Usadel, B. (2019), 'MapMan4: A Refined Protein Classification and Annotation Framework Applicable to Multi-Omics Data Analysis', *Molecular Plant* **12**(6), 879–892.

Walden, N., German, D. A., Wolf, E. M., Kiefer, M., Rigault, P., Huang, X. C., Kiefer, C., Schmickl, R., Franzke, A., Neuffer, B., Mummenhoff, K. & Koch, M. A. (2020), 'Nested whole-genome duplications coincide with diversification and high morphological disparity in Brassicaceae', *Nature Communications* **11**(1).

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., Paux, E., SanMiguel, P. & Schulman, A. H. (2007), 'A unified classification system for eukaryotic transposable elements', *Nature Reviews Genetics 2007 8:12* **8**(12), 973–982

Williams, B. P., Burgess, S. J., Reyna-Llorens, I., Knerova, J., Aubry, S., Stanley, S. & Hibberd, J. M. (2015), 'An untranslated cis-element regulates the accumulation of multiple C4 enzymes in gynandropsis gynandra mesophyll cells', *Plant Cell* **28**(2), 454–465.

Wiludda, C., Schulze, S., Gowik, U., Engelmann, S., Koczor, M., Streubel, M., Bauwe, H. & Westhoff, P. (2012), 'Regulation of the photorespiratory GLDPA gene in C4 Flaveria: An intricate interplay of transcriptional and posttranscriptional processes', *Plant Cell* **24**(1), 137–151.

Wurzinger, B., Nukarinen, E., Nägele, T., Weckwerth, W. & Teige, M. (2018), 'The SnRK1 Kinase as Central Mediator of Energy Signaling between Different Organelles', *Plant Physiology* **176**(2), 1085.

“Geheimnisvoll am lichten Tag,
Lässt sich Natur des Schleiers nicht berauben,
Und was sie deinem Geist nicht offenbaren mag,
Das zwingst du ihr nicht ab mit Hebeln und mit Schrauben.“
Goethe, Faust I

9

Manuscript III

Genetic and epigenetic regulation of *GLDP1* expression in C_3-C_4 intermediate Brassicaceae

Genetic and epigenetic regulation of *GLDP1* expression in C_3-C_4 intermediate Brassicaceae

Sebastian Triesch[1,2], Carina Kontny[1], Rylee Sokoloski[1], Urte Schlüter[1,2], Andreas P.M. Weber[1,2]

1 Institute for Plant Biochemistry, Heinrich Heine University Düsseldorf, Germany

2 Cluster of Excellence on Plant Sciences (CEPLAS)

Author contributions

S.T. conducted or supervised all experiments, analyzed all data and wrote the manuscript. **C.K.** analyzed GUS fusion lines and cloned the M-Box replacement construct under supervision from **S.T.**. **R.S.** cloned the native *DeGLDP1* and M-Box deleted *AtGLDP1* upstream sequences under supervision from **S.T.**. **U.S.** and **A.P.M.W.** conceptualized and supervised the project.

Acknowledgements

The authors thank Gabi Buchmann and Samantha Flachbart for their help in bisulfite library preparation and Thorsten Wachtmeister for sequencing of the bisulfite libraries. Computational infrastructure and support were provided by the Centre for Information and Media Technology at Heinrich Heine University Düsseldorf.

Data availability

All data can be found in an Annotated Research Context Format (ARC) under https://git.nfdi4plants.org/hhu-plant-biochemistry/triesch2024_moricandia_methylation_gldp (private repository as of July 2024).

Abstract

Cell-specific gene expression allows the differentiation of cell types, tissues and organs and is fundamental to all complex traits in higher organisms. The regulation of cell-specific gene expression entails all regulatory features available for an organism, such as selective enhancer or repressor binding, involvement of non-coding RNAs and the regulation of chromatin accessibility histone or DNA modifications.

C_3-C_4 intermediate photosynthesis is one of the complex traits that underwent a sophisticated rewiring of spatial gene activity during the evolution from the ancestral C_3 state. It was previously shown in C_3-C_4 intermediate plants that the expression of key genes underlying this complex trait is confined to the leaf bundle sheath cells (BSC) whereas their expression is found in BSC and mesophyll cells (MC) in closely related species with C_3 photosynthesis. This shift of the expression domain was likely through the recruitment of BSC expression modules to the regulatory upstream sequences of these genes during the evolution of C_3-C_4 photosynthesis. The recruitment or abolishment of regulatory upstream elements is a fundamental feature to gain or lose cell-specific gene expression in the evolution of traits.

The *GLDP* gene, encoding the P-protein of the glycine decarboxylation complex, is a key player in the establishment of a photorespiratory glycine shuttle, an early step in the evolution of C_3-C_s intermediate photosynthesis. It has been shown that the gene is expressed throughout the leaf blade in C_3 plants but selectively only in BSC in C_3-C_4 intermediate plants. Building on the natural variation in the regulatory sequences of *GLDP* genes we here attempt to unravel the mechanism of the specific expression of this gene. We focused on species from the Brassicaceae family where the BSC-specificity of *GLDP* expression was so far linked to the presence of transposable elements in the *GLDP1* upstream sequences of C_3-C_4 intermediate species. We hypothesized that the shift of the *GLDP* expression domain in these species is linked to

the presence and epigenetic modification of these transposable elements. We tested this hypothesis by employing GUS analysis of native and engineered *GLDP1* upstream sequences and epigenetic whole-genome bisulfite sequencing. Our results show a mixed pattern and hint at a potentially more complex regulation of *GLDP1* expression in C_3-C_4 intermediate Brassicaceae.

Introduction

The increased availability and democratization of next-generation sequencing (NGS) techniques has greatly benefited plant research. The optimized speed, cost and capability of NGS and the ever-increasing computational power has advanced the sequencing of a multitude of plant genomes, transcriptomes and epigenetic datasets on whole-plant, tissue and even single-cell level. Especially in non-model and crop plants, the analysis of these datasets has given insights into natural variation of complex photosynthetic traits of interest for agriculture, such as water-use efficient C_4 and CAM photosynthesis (Covshoff *et al.*, 2014; Edwards, 2019).

These large datasets can be explored hypothesis-free, however, conducting genomics with a data-driven approach harbors the risk to assemble large gene regulatory networks and lists of enriched gene ontologies without direct experimental corroboration or generation of functional insights. The dichotomy between hypothesis-free, exploratory research and mechanistic, molecular biology has recently been discussed in computational biology, mathematical modeling and science philosophy literature (Ratti, 2015; Jafari *et al.*, 2021).

The research of water and nitrogen-use efficient forms of photosynthesis like C_4 photosynthesis has benefited substantially from large-data biology. C_4 photosynthesis is a form of photosynthesis with major potential impact on the optimization of yield performance. Since C_4 photosynthesis is highly complex, this study focuses on C_3-C_4 intermediate species that are stable interim stages on the evolutionary path from C_3 to C_4 photosynthesis. In researching the genomics of C_4 and C_3-C_4 intermediate species, the primary focus has been on the analysis of global (Mallmann *et al.*, 2014; Schlüter *et al.*, 2017; Lyu *et al.*, 2021; Dai *et al.*, 2022) or tissue-specific changes in gene expression patterns (Wang *et al.*, 2014; Dai *et al.*, 2022; Singh *et al.*, 2023; Swift *et al.*, 2023) but insights into the regulatory machinery behind species-specific gene regulation significantly lag behind.

Aspects of the underlying functional genomics have been revealed in C_4 and C_3-C_4 intermediate species from the Asteraceae and Brassicaceae families. As these families contain multiple genera with independent evolutionary tracks towards C_4 photosynthesis, they are promising to study the functional genomics underlying C_4 and C_3-C_4 intermediate traits. The differential expression of genes between mesophyll and bundle sheath cells is the central genetic mechanism behind C_4 and C_3-C_4 intermediate photosynthesis. Making use of transcriptomic and genomic data, the *cis*-regulatory elements driving cell-specific expression can be found and their contribution to the establishment of C_4 and C_3-C_4 intermediates can be analyzed.

In C_4 *Flaveria* species, Gowik *et al.* (2004) identified the MEM1 (mesophyll expression module 1) element in the upstream region of the mesophyll-specific *ppcA* gene. This element was sufficient to drive mesophyll-specific expression of the β-glucuronidase (*GUS*) gene in transgenic *Arabidopsis thaliana* lines. The MEM1 element consists of two regions, including a CACT tetranucleotide that was shown to be present selectively in C_4 and C_3-C_4 intermediate homologs of the *ppcA* upstream region. An upstream element with high homology to the MEM1 element was identified in the upstream region of the mesophyll-specific *Flaveria* C_4 isoform of CARBONIC ANHYDRASE (*CA3*).

This MEM1-like element was also shown to drive mesophyll-specific expression in the transgenic *A. thaliana* GUS fusion system (Gowik *et al.*, 2017). In the upstream region of the gene encoding the glycine decarboxylase complex (GDC) T-protein (*GLDT*), the insertion of a transposable element (TE) was observed selectively in C_4 and C_3-C_4 intermediate *Flaveria* species, in which the *GLDT* gene is selectively expressed in bundle sheath cells (BSC; Emmerling (2018)). Upon deletion of the TE, the upstream sequence lost its ability to mediate BSC-specific *GUS* expression in transgenic *A. thaliana* lines. The author assumed that in C_4 and C_3-C_4 intermediate species, the TE in the *GLDT* upstream region acts as a spacer between proximal and distal promoter elements and causes BSC-specificity.

As illustrated in the latter example, the mechanisms underlying differential expression of GDC genes have already received attention in research on C_3-C_4 intermediate plants. It is assumed that the expression domain shift of GDC genes to the BSC is one of the first evolutionary steps towards C_3-C_4 intermediate photosynthesis. Its restriction to the BSC seems to be convergent between C_3-C_4 intermediate species across various plant families. The shift of the GDC to the BSC, accompanied with anatomical adjustments, seems to be sufficient to install a photorespiratory glycine shuttle that decreases the oxygenation reaction of Rubisco and improves photosynthetic efficiency. (Sage, 2004). The GDC is composed of four subunits, the H-, L-, T- and P-protein. Early studies reported that the BSC-preferential activity of the GDC in C_3-C_4 intermediate plants is largely due to the BSC-specific expression of *GLDP* genes, encoding the P-protein subunit of the GDC (Rawsthorne *et al.*, 1988a; Schulze *et al.*, 2016).

In the Brassicaceae, the *GLDP2* copy was shown to be absent in the Brassiceae tribe which contains all known C_3-C_4 intermediate Brassicaceae species (Schlüter *et al.*, 2017) and we found the *GLDP1* gene to be differentially partitioned between MC and

BSC in C_3-C_4 intermediate *Moricandia arvensis* (Triesch *et al.* (2024), Chapter 7). In a first comparative study, Adwy *et al.* (2015) aligned *GLDP1* upstream sequences across eight Brassicaceae species. They identified two conserved sequence elements in these regions, which were named based on the proposed function of the elements: the first element (M-Box) resides around 500-1000 nt upstream of the *GLDP1* start codon, in the promoter region of that gene. The second element (V-Box) was found in the proximal *GLDP1* promoter, approximately 200-500 nt upstream. When removing the M-Box in a promoter truncation approach, the truncated *A. thaliana* promoter fused to the *GUS* report gene showed vein specific GUS staining, whereas the full-length promoter showed GUS staining in the whole leaf blade. Upon further truncation, which involved removing the V-Box, the promoter failed to induce any *GUS* expression. The authors concluded that the M-Box contains transcription factor binding sites (TFBS) for mesophyll-expression, whereas the V-Box contains TFBS that drive expression in the BSC. In a more recent study, we revealed that the BSC-expression conveying element in the proximal *GLDP1* promoter is a dual MYC/MYB TFBS outside the V-Box (Dickinson *et al.*, 2023). This MYC/MYB element was also found to drive the expression of BSC-specific genes from glucosinolate biosynthesis in the *A. thaliana* BSC (Dickinson *et al.*, 2020). The recruitment of promoter element from different genes involved in different pathways is a good example for the rewiring of gene regulatory networks in the evolution of C_4 photosynthesis.

In the M-Box region, the sequence element driving *GLDP1* expression in the MC was not identified to date. Using the yeast-one-hybrid system, a GATA5 transcription factor was shown to bind the M-Box region, but no further functional corroboration studies were conducted (Adwy, 2018). It is so far unknown which TFBS the M-Box region might entail or to what extend it mediates MC expression of *GLDP1*. In *A. thaliana*, the M-Box sequence was deleted using the CRISPR-Cas9 technique (Hahn, 2018). Elevated levels of glycine could be detected in leaves of these M-Box deletion mutant lines, suggesting a disruption of GDC activity, abundance or localization, leading to glycine accumulation (unpublished work).

Screening a panel of C_3 and C_3-C_4 intermediate Brassicaceae, Adwy *et al.* (2019) suggested a loss of the M-Box from the *GLDP1* promoters selectively in C_3-C_4 intermediate species. However, we could show that the M-Box is in fact present in all analyzed species, but shifted upstream by TE insertions in C_3-C_4 intermediates. Interestingly, this shift can be observed in species from four different evolutionary origins of C_3-C_4 intermediate photosynthesis by four different types of transposons (Triesch *et al.* (2024), Chapter 8). In *M. arvensis*, *M. nitens* and *M. suffruticosa*, a miniature inverted-repeat TE (MITE) was identified downstream of the M-Box, in contrast to the closely related C_3 plant *M. moricandioides*. In *Diplotaxis tenuifolia*, a TE from a different TE class inserted into the *GLDP1* promoter downstream of the M-Box. In the C_3-C_4 intermediate *Hirschfeldia incana* HIR3 (otherwise referred to as *Sinapis pubescens*, Guerreiro *et al.* (2023)), a *Helitron* TE could be detected downstream of the M-Box, that was absent in the closely related C_3 plant *H. incana* HIR1. Whereas in these examples the M-Box was shifted around 1,000-2,000 nt further upstream in the C_3-C_4 intermediate genomes compared to the C_3 counterparts, in *Brassica gravinae* we found a large upstream shift of the M-Box of around 4,000 nt compared to the C_3 genomes. We assumed that this shift was due to the nested insertion of multiple TEs (Triesch *et al.* (2024), Chapter 8). Interestingly, the TE insertion in the *B. gravinae GLDP1* upstream region caused a shift of both the M-Box and the V-Box, but not the MYC. This confirmed the role of MYC/MYB TFBS in BSC gene expression, rather than the previously proposed V-Box. One exception was *D. erucoides* that exhibits C_3-C_4 intermediate characteristics (Schlüter *et al.*, 2023) but did not show an M-Box shift (Triesch *et al.* (2024), Chapter 8).

These TE insertions and the associated shift of the M-Box correlated highly to the presence of C_3-C_4 intermediate characteristics in the analyzed set of Brassicaceae species. Thus, we hypothesized that the TE insertion is the causal factor underlying the differential expression of *GLDP1* between C_3 and C_3-C_4 intermediate Brassicaceae. We speculated that the TE functions as a spacer that elongates the distance to between the M-Box and the core promoter and thus precludes biding of transcription factors M-Box. Alternatively, we hypothesized that epigenetic modification in the TE such as DNA methylation has an impact on *GLDP1* regulation (Triesch *et al.* (2024), Chapter 8). However, there are still no supporting experiments explaining how and to which extend these factors influence *GLDP1* expression.

In this study, we examined the influence of epigenetic modification and structural variation in the *GLDP1* upstream region on *GLDP1* expression patterns. To gain a solid understanding of the architecture of the functional *GLDP1* promoter elements, we focused on the *GLDP1* upstream regions of C_3 *A. thaliana* and *M. moricandioides* as well as C_3-C_4 intermediate *M. arvensis* and *D. erucoides*. Performing methylation sequencing, we sought to show the epigenetic impact of DNA methylation in the TE region on *GLDP1* regulation and generated various promoter truncation DNA constructs to reveal the function of the M-Box sequence and the TE insertions.

Material and Methods

Genomes

Genomic data for promoter analysis, whole-genome bisulfite sequencing read mapping and primer design was obtained from Guerreiro *et al.* (2023). All structural gene model annotations were created using *Helixer* (Stiehler *et al.*, 2020; Holst *et al.*, 2023) in Triesch *et al.* (2024). Alignments were performed using *Seaview* 5.0.5 (Gouy *et al.*, 2010) and visualized using *pyMSAviz* 0.4.2.

Whole Genome Bisulfite Sequencing and Data Analysis

Genomic DNA of mature *M. arvensis* and *M. moricandioides* leaves was isolated using the *QIAGEN DNeasy Plant Kit* in three biological replicates per species. 400 ng of the genomic DNA was sheared using a *COVARIS M220 NGS sample processor* using the following settings: temperature 12 °C, treatment time 10 s, peak inc. power 70, duty factor 20, cycles per burst 1000, repetitions 10. The resulting DNA fragment distributions were analyzed on the *AGILENT 2100 Bioanalyzer*. The sheared DNA was processed using the *NEBNext® Enzymatic Methyl-seq Kit* following the "standard insert sized" protocol. This included the addition of unmethylated lambda phage DNA and fully methylated pUC19 plasmid DNA to all samples as spike-in controls. The resulting NGS libraries were sequenced on the *Illumina HiSeq 2000* platform.

Raw read quality was verified using *fastqc* 0.11.5 and *multiqc* 1.11. Reads were trimmed using *cutadapt* 3.5 using the following parameters: -a AGATCGGAAGAGCACACGTCTGAAC -g AGATCGGAAGAGCGTCGTGTAGGGA --length 76 --minimum-length 25 --quality-cutoff 5. The processed reads were mapped and the methylation levels were quantified using *Bismark* 0.23.1 (Krueger *et al.*, 2011). All downstream analyses were performed using *python* 3.8.8.

Cloning

A binary vector for the easy and modular cloning of upstream regions and transcriptional fusion to a *GUS* reporter was created by inserting the GUS-encoding gene including a *nos* terminator into the binary pB415 vector using *Sfi*I. A *Nco*I restriction site was added to the start site of the *GUS* gene. This allowed upstream regions of interest to be inserted in full length, inducing their 5' UTR up to their native start codon. The resulting vector (cST6) was designed to accept upstream regions using *Pst*I and *Nco*I.

Full length and engineered *GLDP1* upstream sequence variants were transcriptionally fused to the *GUS* reporter gene in the cST6 vector. Transgenic *A. thaliana* Col-0 plants were created using the floral dip method (Feldmann *et al.*, 1987). T1 lines were selected on ½ strength Murashige-Skoog medium agar plates supplemented with kanamycin.

GUS staining

Young leaves of transgenic *A. thaliana* plants were cut and placed in GUS staining solution (5 mM $K_3[Fe(CN)_6]$, 5 mM $K_4[Fe(CN)_6]$ x $3H_2O$, 0.1 M $NaPO_4$ buffer (pH 6.7), 0.5 mg/mL X-Gluc) before applying vacuum three times for 5 min. The infiltrated leaves in the GUS staining solution were incubated overnight at 37 °C. After the overnight staining, the GUS staining solution was removed and the leaves were placed into GUS fixation solution (75 % (v/v) ethanol, 25 % (v/v) acetic acid) and incubated for 10 min at 65 °C. The fixation solution was removed and the leaves were destained twice in 80 % (v/v) ethanol.

Results

Multiome analysis of the *GLDP1* upstream region

We initiated our investigation of the genetic and epigenetic architecture of the *GLDP1* upstream region by integrating public and unpublished NGS datasets and machine learning predictions for gene models and the chromatin state of the *GLDP1* region. Underlying these investigations were the genome assemblies of C_3 and C_3-C_4 intermediate species (Guerreiro *et al.*, 2023) which were used to map NGS datasets and as inputs for *Helixer* and *Predmoter*. Using multiple sequence alignment (MSA) of the *GLDP1* upstream region across 15 Brassicaceae species, we found the two previously reported conserved regions. The MYC/MYB TFBS was found in the proximal promoter of the *GLDP1* gene. The M-Box region located about 1,000 nt upstream of the *GLDP1* start codon in C_3 plants and approximately 2,000 nt upstream of the *GLDP1* start codon in genomes C_3-C_4 intermediate plants (Supp. Fig 1).

Two enhancers characterized in self-transcribing active regulatory region sequencing (STARR-seq) experiments for *A. thaliana* (Jores, unpublished) aligned to the *Moricandia GLDP1* upstream region (Fig. 1). One of the enhancer STARR-seq sequences was found in the highly conserved M-Box region if *M. arvensis* and *M. moricandioides*. In the STARR-seq dataset, this enhancer sequence showed increased expression in light compared to darkness and no orientation-specificity. Another STARR-seq enhancer sequence mapped to the MYC/MYB TFBS in both *Moricandia* species. In the STARR-seq experiment, this sequence showed overall lower expression strength and no expression once the sequence was reverted.

As a further line of evidence for the enhancer characteristics of the M-Box region, we used *Predmoter* (Kindel *et al.*, 2024) to predict the chromatin state of the *GLDP1* upstream region. Within both *Moricandia* species, we observed significantly increased predicted ATAC-seq and ChiP-seq coverage in the region of the *GLDP1* transcriptional start site and first exon for both species. In the M-Box only slight increases in predicted ATAC-seq coverage could be found, which were only marginally increased over the background noise (Fig. 1).

Whole-genome DNA methylation sequencing of *Moricandia* species

Since TEs are frequently targeted by DNA methylation, we sought to experimentally quantify the methylation density in the *GLDP1* locus using methylation sequencing. To this end, we generated bulk whole-genome bisulfite sequencing libraries for mature *M. moricandioides* and *M. arvensis* leaf tissue. We validated the quality of the data by showing a strong decrease in cytosine methylation in 5,000 randomly sampled transcription start sites (TSS enrichment) as well as a strong methylation in intron and intergenic regions as well as in TEs (Supp. Fig. 2 & 3).

In the *GLDP1* upstream region of the C_3-C_4 intermediate *M. arvensis*, we found high methylation levels in the region of the TE insertion that we had previously identified (Triesch *et al.* (2024), Chapter 8). The cytosine methylation in this area was observed mainly in the CG and CHG (H corresponds to A, C or T) context, however it was less pronounced in the CHH context. Interestingly, we found the cytosine methylation to extend slightly beyond the borders of the MITE (Supp. Fig. 4). This spread partially methylated cytosines in the M-Box region specifically in the GATA TFBS identified by Adwy (2018). In contrast, the *GLDP1* upstream region containing the M-Box was completely unmethylated in the C_3 species *M. moricandioides*.

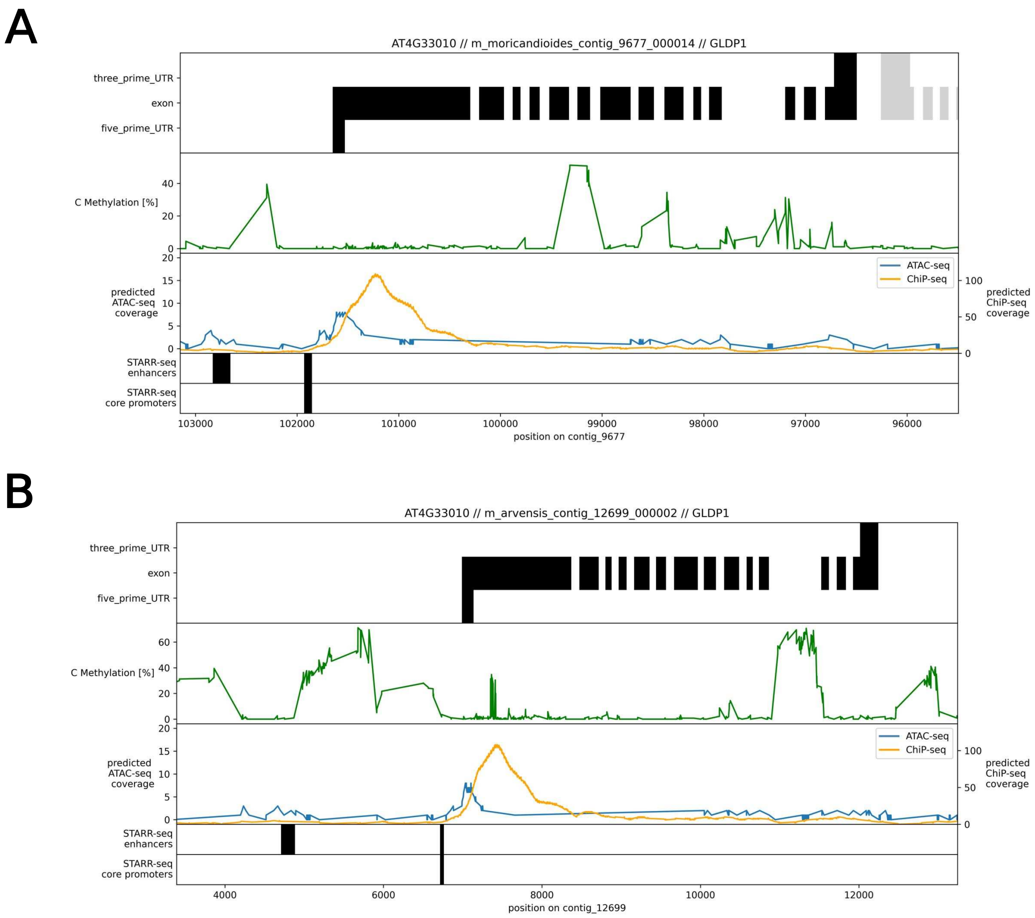


Fig. 1: **Multiome integration plots for the *Moricandia moricandioides* (A) and *M. arvensis* (B) *GLDP1* upstream region**. The upper panel of each plot shows gene model predictions from *Helixer* (Schwacke *et al.*, 2019; Holst *et al.*, 2023), the second panel shows cytosine DNA methylation in the CG context. The third panel shows ATAC-seq and ChIP-seq predictions using *predmoter* (Kindel *et al.*, 2024) and the lowermost two panels show BLAST hits for enhancer and core promoter STARR-seq experiments from *A. thaliana* (Jores, unpublished).

Experimental promoter analysis *via GUS* staining

Based on the structural and epigenetic variation found between the *M. arvensis* and *M. moricandioides GLDP1* upstream region, we hypothesized that the shift of the M-Box and/or its DNA methylation is the functional reason for the shift of the *GLDP1* expression domain to the BSC in *M. arvensis*. To corroborate this hypothesis, we generated multiple transgenic *A. thaliana* lines, heterogously expressing the *GUS* reporter gene under the control of native or engineered *GLDP1* promoter variants. In all transgenic *A. thaliana* T1 lines, the staining patterns were analyzed in leaves of different ages.

First, we reanalyzed the native *GLDP1* upstream regions from *M. arvensis* and *M. moricandioides* which have been previously analyzed in Lin (2020). *A. thaliana* lines expressing *GUS* under the native *MaGLDP1* promoter from *M. arvensis* (line ST1) showed vein-specific staining. Lines with *GUS* expression controlled by the *MmGLDP1* upstream region (line ST2) showed staining in the whole leaf blade (Fig. 2, Supp, Fig. 6).

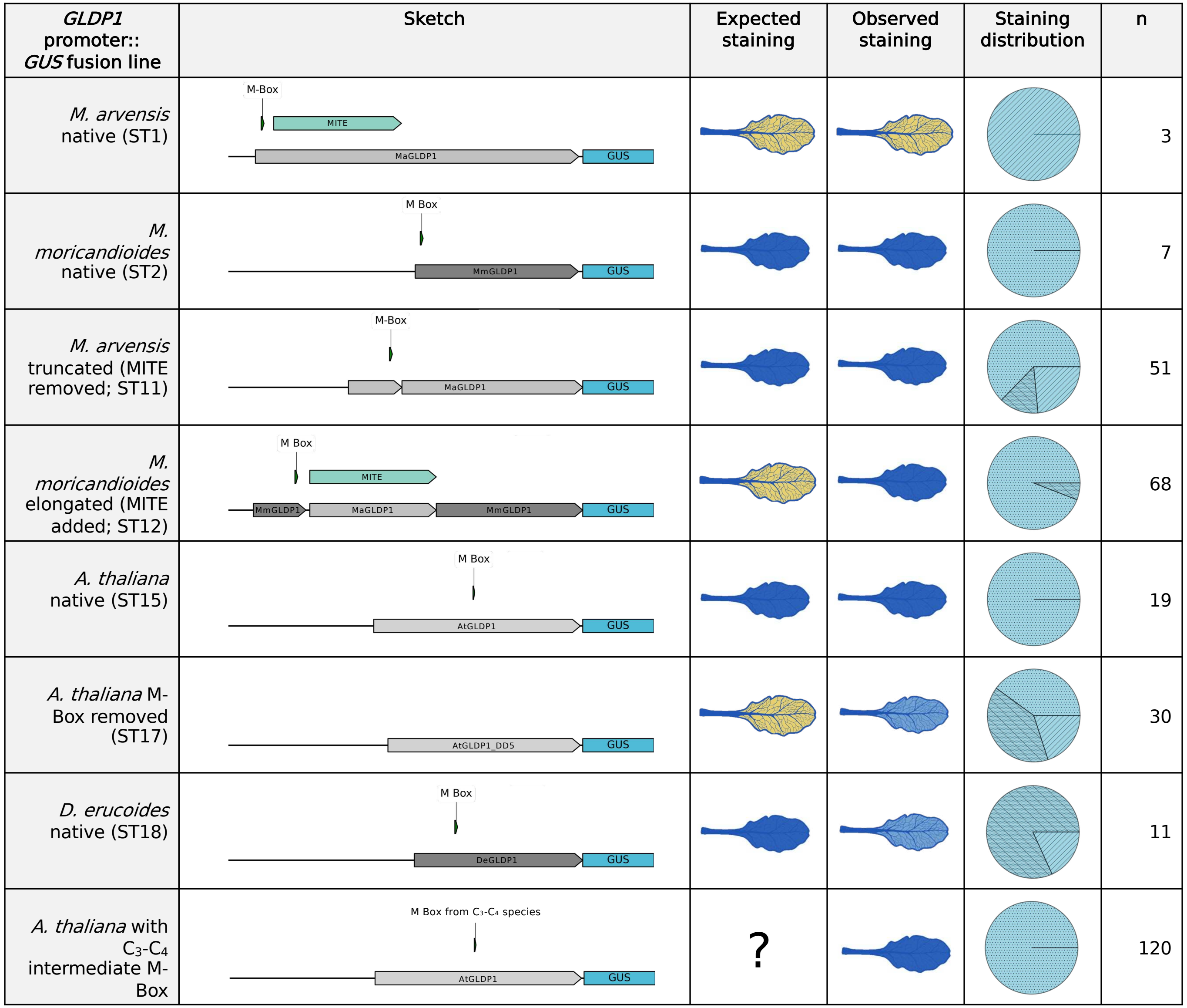
Next, we analyzed the influence of the TE insertion in the C_3-C_4 intermediate *GLDP1* upstream region. We created a truncated version of the *MaGLDP1* upstream region (line ST11), excising the TE at the target site duplications identified in Triesch *et al.* (2024). Thereby, the spacing in the *MmGLDP1* upstream region was mimicked, with the M-Box closer to the *GLDP1* start site. The analysis of 51 ST11 T1 lines revealed an inconsistent pattern. Around 25 % of the lines showed vein-specific GUS patterns, 60 % showed ubiquitous staining and the remaining lines showed a mixed pattern with prominently stained veins on a stained background leaf blade.

Additionally, we identified the homologous TE insertion site in the in the *MmGLDP1* upstream region and integrated the TE from *MaGLDP1* at this position (line ST12), thus mimicking the M-Box spacing in C_3-C_4 intermediates. Fusion of this elongated *MmGLDP1* promoter to *GUS* in the *A. thaliana* lines showed ubiquitous GUS staining in the whole leaf in 94 % of the T1 lines, whereas 6 % of the T1 lines showed veins with more prominent staining.

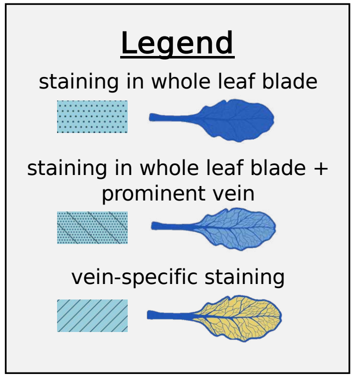
Exploiting more natural diversity in the *GLDP1* region, we also used the -1200 bp upstream region from *D. erucoides*, a plant showing C_3-C_4 intermediate characteristics but no TE insertion in the *GLDP1* upstream region. Expression of *GUS* controlled by the *DeGLDP1* upstream sequence (line ST18) resulted in prominently stained veins, however significant background staining in the mesophyll was visible. Around 20 % of the analyzed lines T1 showed clear vein-specific GUS staining patterns.

Furthermore, we sought to understand the impact of natural single nucleotide variation in the M-Box region on *GLDP1* expression patterns. To this end, we focused on the *GLDP1* upstream region from *A. thaliana*. GUS staining under the control of the native *AtGLDP1* upstream region was detectable in the whole leaf blade in all analyzed lines (ST15). We removed a 100 bp sequence stretch containing the M-Box and fused the shortened *AtGLDP1* promoter to the *GUS* reporter, creating an *A. thaliana* line expressing *GUS* controlled by an M-Box-free *AtGLDP1* upstream sequence (ST17). Here, the pattern was highly inconsistent, with roughly a third of the lines exhibiting GUS staining in the whole leaf, in more prominent veins against a stained mesophyll background or vein-specifically.

In this truncated M-Box-free *AtGLDP1* upstream sequence, we inserted the M-Box regions of several species individually. Specifically, we integrated the homologous M-Box regions from *M. moricandioides* (C_3, line CK1), *M. arvensis* (C_3-C_4 intermediate; line CK2), *D. erucoides* (C_3-C_4 intermediate, line ST27), *D. tenuifolia* (C_3-C_4 intermediate, line CK5) and the two *D. muralis GLDP1* copies, originating from the C_3 (line CK3) and C_3-C_4 intermediate (line CK4) parents. All lines expressing *GUS* under control of the *AtGLDP1* upstream regions with the substituted M-Box variants showed staining in the whole leaf blade. None of the T1 lines showed increased staining in veins or even vein-specific staining.



GLDP1 promoter:: *GUS* fusion line	Sketch	Expected staining	Observed staining	Staining distribution	n
M. arvensis native (ST1)	M-Box MITE MaGLDP1 GUS				3
M. moricandioides native (ST2)	M Box MmGLDP1 GUS				7
M. arvensis truncated (MITE removed; ST11)	M-Box MaGLDP1 GUS				51
M. moricandioides elongated (MITE added; ST12)	M Box MITE MmGLDP1 MaGLDP1 MmGLDP1 GUS				68
A. thaliana native (ST15)	M Box AtGLDP1 GUS				19
A. thaliana M-Box removed (ST17)	AtGLDP1_DD5 GUS				30
D. erucoides native (ST18)	M Box DeGLDP1 GUS				11
A. thaliana with C_3-C_4 intermediate M-Box	M Box from C_3-C_4 species AtGLDP1 GUS	?			120



Legend

staining in whole leaf blade

staining in whole leaf blade + prominent vein

vein-specific staining

Fig. 2: Overview of the GUS staining experiments for native and engineered *GLDP1* upstream sequences. The first column indicates the used *GLDP1* upstream sequence fused to the *GUS* reporter gene. The second column depicts a schematic overview of the cloned *GLDP1* upstream region. The second and third column illustrate the expected and observed GUS staining patterns in three categories: staining in the whole leaf blade (full blue schematic leaf), staining in the whole leaf blade with more prominently stained veins (light blue leaf schematic with thicker veins) or vein-specific staining (yellow leaf schematic with blue veins). The pie charts in the fifth column illustrate how many leaves from the T1 generation show staining within these defined categories. The number in column six is the total number of tested leaves from individual T1 lines.

Discussion

The differential localization of proteins of the GDC has been one of the first findings in the molecular biology research of C_3-C_4 intermediate species. In the 1980s, the GDC P-protein, catalyzing the glycine decarboxylation step, was shown to localize specifically to the *M. arvensis* BSC mitochondria (Hylton *et al.*, 1988; Rawsthorne *et al.*, 1988b; Rawsthorne *et al.*, 1988a; Rawsthorne, 1992; Morgan *et al.*, 1993; Monson *et al.*, 2000). Since then, multiple studies revolved around the functional genetics of BSC-specificity, especially for the genes encoding the GDC proteins.

Here, we studied the function of the *GLDP1* upstream region in conferring BSC-specific gene expression. Previous experiments could narrow down the functional parts of the *GLDP1* upstream region to two conserved elements, the M-Box (Adwy *et al.*, 2015) and a dual MYC/MYB binding motif in the proximal promoter (Dickinson *et al.*, 2023). The M-Box region, the potential regulator of mesophyll expression, was reported to be absent in C_3-C_4 intermediate Brassicaceae, explaining the BSC-specific expression of *GLDP1*. However, we found the M-Box to be present in a large panel of C_3 and C_3-C_4 intermediate Brassicaceae, but shifted upstream by a TE insertion in most C_3-C_4 intermediate genomes. This contradicts previous studies and assumptions about the mechanisms of the *GLDP1* expression domain shift.

The shift of the M-Box correlates highly with the lowered carbon compensation point of the C_3-C_4 intermediate Brassicaceae species, potentially underlying the shift of *GLDP1* expression to the BSC (Triesch *et al.* (2024), Chapter 8). We previously showed that single-nucleotide exchanges in the M-Box region do not correlate with the presence of a C_3-C_4 intermediate lifestyle, which strengthens the hypothesis that the TE insertion is causal for the shift of the *GLDP1* expression domain to the BSC. In this study, we sought to corroborate this hypothesis by detailed promoter analysis and promoter truncation experiments.

Promoter truncation experiments were conducted by fusion of the native and altered *GLDP1* upstream sequences to the *GUS* reporter gene and analysis of transgenic *A. thaliana* T1 lines transformed with this assembly. Using this experimental setup, analysis of the native C_3 and C_3-C_4 intermediate *GLDP1* upstream region from *M. moricandioides* and *M. arvensis* showed the expected results, confirming that the C_3 *MmGLDP1* region confers *GUS* expression in MC and BSC, whereas the C_3-C_4 intermediate *MaGLDP1* upstream region including the MITE restricted *GUS* expression to the veins. By swapping the MITE from the *MaGLDP1* upstream region to the *MmGLDP1* upstream region, we expected to see a reversed pattern, since the *MmGLDP1* M-Box was now shifted significantly further upstream and vice versa. However, we saw a highly inconsistent pattern between the analyzed T1 lines. Testing the truncated *MaGLDP1* upstream construct, where the MITE was excised, the majority of the lines showed GUS staining in the whole leaf blade. However, a significant proportion of the analyzed lines showed a strong staining along the veins or even vein-specific staining (Fig. 2 & Supp. Fig. 6). This leads to the conclusion that the removal of the TE alone is not sufficient to abolish BSC-specificity conferred by the *GLDP1* upstream region. Insertion of the MITE to the *MmGLDP1* upstream region was expected to result in vein-specific staining due to an upstream shift of the M-Box. However, largely ubiquitous staining was observed in the leaves of transgenic *A. thaliana* T1 leaves transformed with this construct. Only a few lines showed vein-preferential staining, but with a strong MC background (Fig. 2 & Supp. Fig. 6). The analysis of the engineered *Moricandia GLDP1* upstream region suggested an influence, but not a sole effect, of the TE insertion on GUS staining patterns.

We also analyzed the *GLDP1* upstream region from *D. erucoides*, which lacks a TE insertion and features the M-Box positioned in a C_3-like promoter architecture. We expected the GUS staining directed by the *DeGLDP1* upstream region either in the whole leaf blade, as suggested by the short distance between the M-Box and the start codon, or along the veins, as suggested by the C_3-C_4 intermediate characteristics of *D. erucoides*. However the observed GUS staining showed a remarkable pattern, with prominently stained veins standing out against a stained MC background (Fig. 2 & Supp. Fig. 6). This pattern could hint at variation in the *GLDP1* upstream region beyond the TE insertion in other C_3-C_4 intermediate Brassicaceae. It is unclear whether the causal variation manifests as single nucleotide polymorphisms (SNPs) or in other means of regulation. As a representative from an individual evolutionary origin of C_3-C_4 intermediate photosynthesis, it is well conceivable that *D. erucoides* uses other means to establish a glycine shuttle than the predominantly BSC-specific expression of *GLDP1* or other characteristics of the C_2 cycle, potentially also with other shuttle molecules involved. In metabolomic analyses for example, *D. erucoides* clustered separately from the other C_3-C_4 intermediate species, indicating a different metabolic makeup of photorespiratory shuttles in this plant (Schlüter *et al.*, 2023).

To test the influence of SNPs within the M-Box region on *GLDP1* expression, we examined M-Box sequences of different species in the *A. thaliana AtGLDP1* upstream sequence. Analysis of the M-Box free *AtGLDP1* sequence showed a similar GUS staining pattern as derived by the *DeGLDP1* sequence. Prominently stained veins were visible, but a significant staining of mesophyll tissue was also observed. However, the patterns were highly inconsistent, with observations of both whole leaf blade and vein-specific GUS staining. This indicates that the M-Box sequence might not be the only determinant of MC-specific expression but suggests other additional factors likely required. This is in contrast to previous promoter truncation experiments (Adwy *et al.*, 2015) which observed an abolished MC-specific expression consequent of a truncated M-Box.

When we replaced the native M-Box in the *AtGLDP1* upstream sequence with those of other C_3 and C_3-C_4 intermediate species, ubiquitous GUS staining was observed. This indicates that SNPs in the M-Box sequence are not causal for the spatial expression shift in C_3-C_4 intermediates.

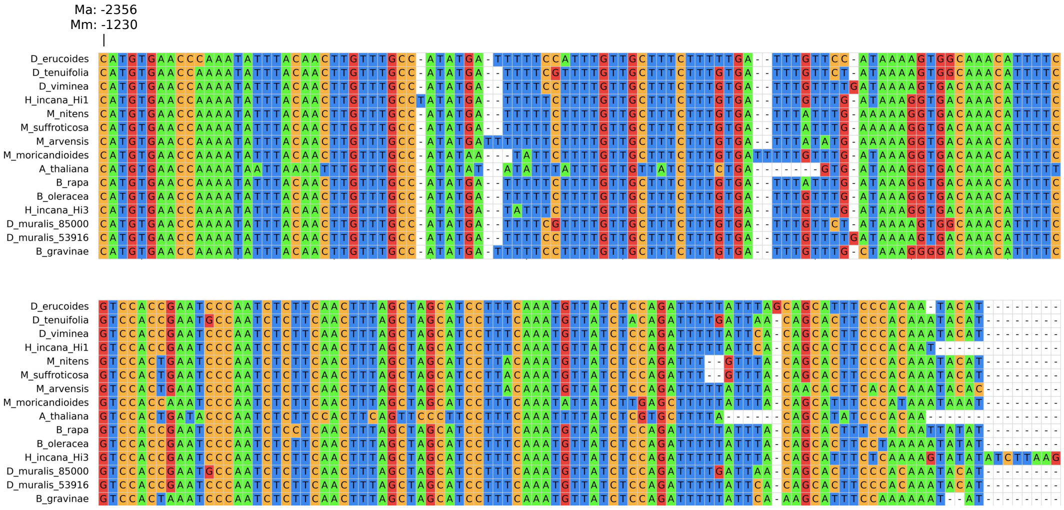
The highly inconsistent staining patterns in the analyzed T1 may be due to the genomic surroundings of the random T-DNA insertion carrying the *GLDP1* upstream sequences fused to *GUS*. We tried to minimize this effect by analyzing a high number of individual T1 lines. However, in the constructs using the native upstream regions, GUS staining is remarkable consistent between individual T1 lines. We therefore hypothesized that *GLDP1* also underlies means of regulation that are beyond classical genetics.

Following up on this observation, we employed whole-genome bisulfite sequencing to quantify levels of the 5-methylcytosine DNA modification in *M. moricandioides* and *M. arvensis*. Here, we observed significant methylation of the MITE in the *MaGLDP1* upstream region (Fig. 1). Interestingly, we also discovered that this methylation spreads beyond the TE borders, even extending into the M-Box sequence. This leaves room to speculate about a further epigenetic means of gene regulation affecting *GLDP1*. It is questionable whether this methylation alters the chromatin structure or the accessibility of transcription factor binding motifs to *trans*-elements. The DNA methylation data obtained in this study is bulk data for all leaf cell types. However, there can be cell-type specific methylation patterns that also depend on the availability of C1 groups for methylation (Crider *et al.*, 2012; Groth *et al.*, 2016). Photorespiration is tightly interconnected to the folate cycle that provides these C1 groups (Shi *et al.*,

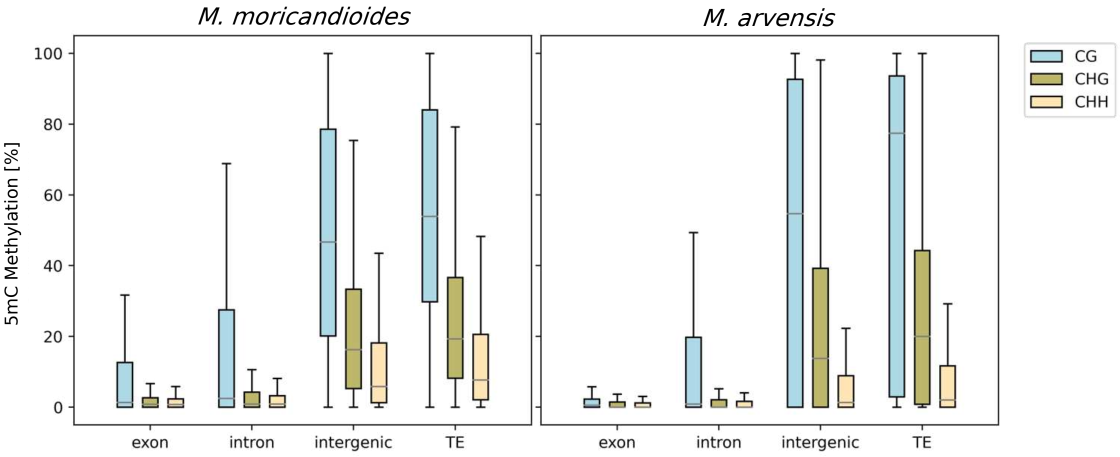
2021) and variation in the availability of C1 groups in the C_3-C_4 intermediate BSC might lead to cell-type dependent methylation of the TE. The M-Box sequence and the epigenetic decoration of the element might not be faithfully represented by GUS staining using the C_3 model plant *A. thaliana*. Methods to quantify the accessibility of the M-Box sequence or the MITE methylation with single-cell resolution exist but are still technically cumbersome (Lu *et al.*, 2016; Grandi *et al.*, 2022).

A more comprehensive analysis of the *GLDP1* upstream region will reveal a more precise understanding of its function in spatially differential *GLDP1* regulation. This can entail additional promoter truncation experiments to identify relevant promoter elements mediating cell-specific expression. The analysis of the epigenetic contribution to *GLDP1* regulation can be continued by analyzing GUS lines in methylation-deficient background lines or in various environmental conditions. Narrowing down the functional regions for BSC-specific expression will also expand the toolkit of regulatory sequences for applications in synthetic biology.

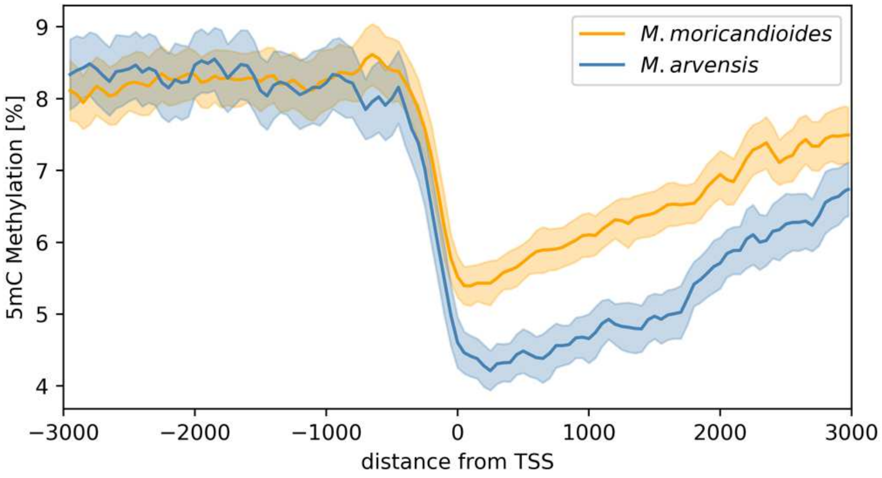
Supplementary Material



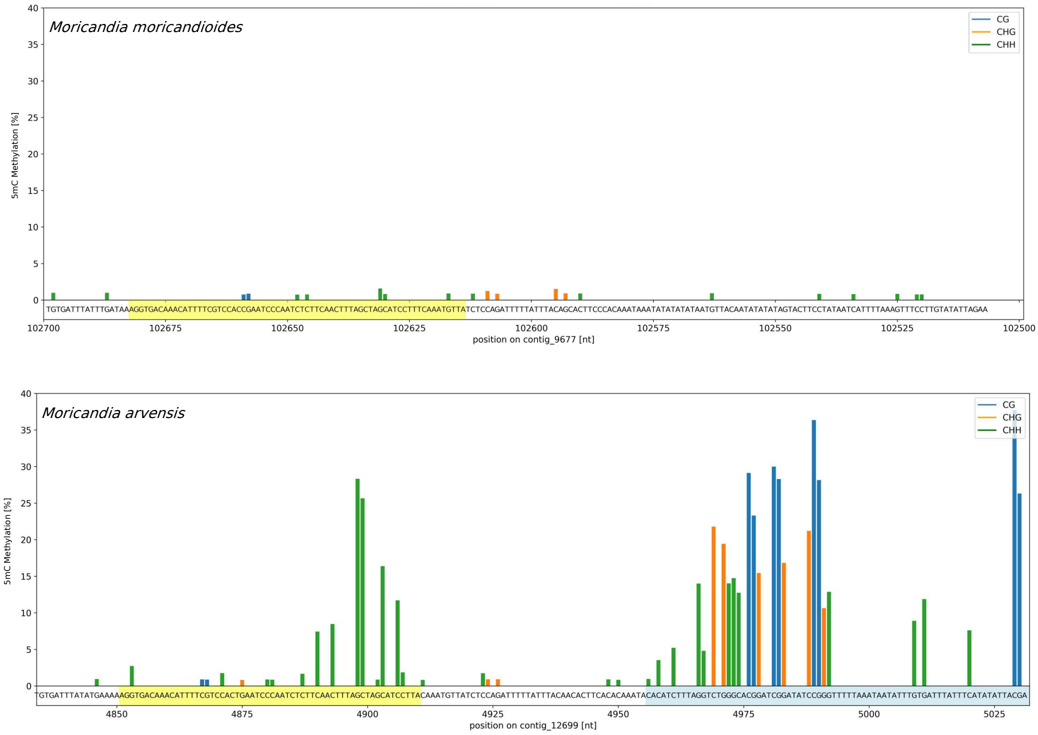
Supp. Fig. 1: Partial multiple sequence alignment of the M-Box region 15 *GLDP1* upstream sequences. The reference position upstream of the *GLDP1* start codon is referenced for *Moricandia arvensis* (Ma) and *M. moricandioides* (Mm).



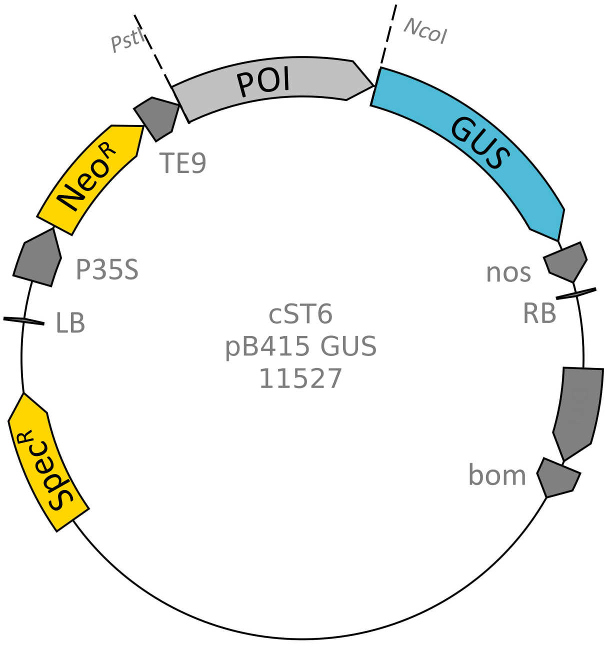
Supp. Fig. 2: 5mC Cytosine methylation levels for all cytosines in the reference genomes for *Moricandia arvensis* and *M. moricandioides*. For each cytosine in the genomes, its position in an exon or intron or between genes was determined using *Helixer*. Transposable elements (TEs) were annotated using EDTA. All annotations were taken from Triesch *et al.* (2024).



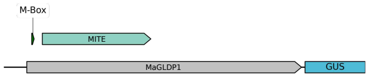
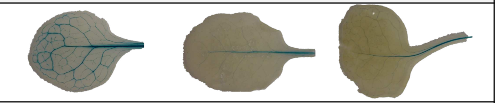
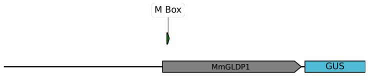
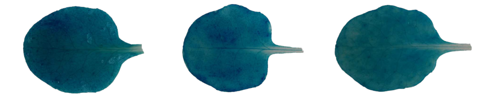
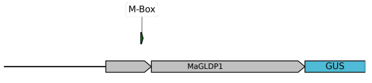
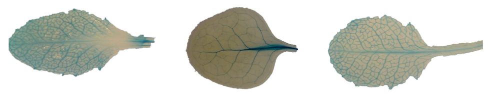
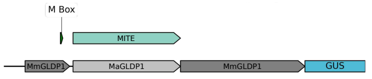
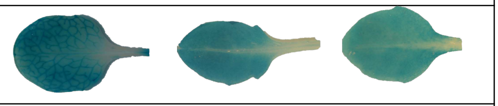
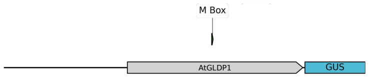
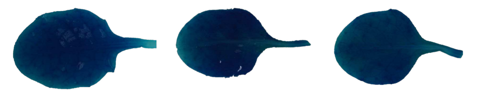
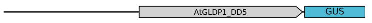
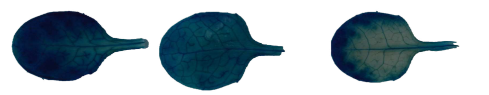
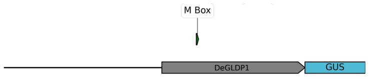
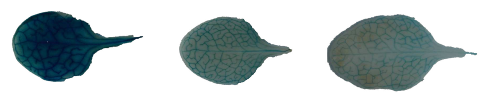
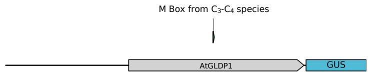
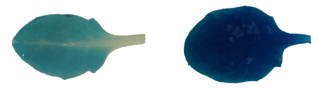
Supp. Fig. 3: 5mC Cytosine methylation levels for all cytosines around the transcription start sites (TSS) of 5000 random genes for *Moricandia arvensis* and *M. moricandioides*. The mean methylation levels were taken for all cytosines in a sliding window of 50 nt.



Supp. Fig. 4: 5mC Cytosine methylation levels for all in the M-Box region in the *GLDP1* upstream regions for *Moricandia arvensis* and *M. moricandioides*. The M-Box region was highlighted in yellow, the transposable element in *M. arvensis* was highlighted in blue.



Supp. Fig. 5: Illustration of the cST6 vector that allows the easy and fast ligation of upstream regions of interest (POI) to the GUS gene. Ligation with *Nco*I allows to retain the native spacing between the native and *GUS* start codon. Abbreviations: Spec^R: spectinomycine resistance gene, LB/RB: left/right T-DNA border; Neo^R: kanamycine resistance gene; P35S: 35S promoter from the cauliflower mosaic virus; TE9/nos: TE9 and nos terminator sequences; bom: origin of replication

GLDP1 promoter:: *GUS* fusion line	Sketch	Selected GUS stained leaf images
M. arvensis native (ST1)		
M. moricandioides native (ST2)		
M. arvensis truncated (MITE removed; ST11)		
M. moricandioides elongated (MITE added; ST12)		
A. thaliana native (ST15)		
A. thaliana M-Box removed (ST17)		
D. erucoides native (ST18)		
A. thaliana with C_3-C_4 intermediate M-Box		

Supp. Fig. 6: Overview over GUS staining patterns in leaves of transgenic *Arabidopsis thaliana* T1 lines expressing *GUS* under control of native and engineered *GLDP1* upstream regions. The illustrations show the *GLDP1* upstream architecture in the T-DNA of the transgenic lines. The leave images are selections with representative GUS staining patterns for the respective lines.

Literature

Adwy, W. (2018) *Identification and characterization of a frequent genetic alteration toward the evolution of C2-photosynthesis in the genus Moricandia.* PhD Thesis, Gottfried Wilhelm Leibniz Universität Hannover.

Adwy, W., Laxa, M. and Peterhansel, C. (2015) 'A simple mechanism for the establishment of C2-specific gene expression in Brassicaceae', *Plant Journal*, 84(6), pp. 1231-1238.

Adwy, W., Schlüter, U., Papenbrock, J., Peterhansel, C. and Offermann, S. (2019) 'Loss of the M-box from the glycine decarboxylase P-subunit promoter in C2 Moricandia species', *Plant Gene*, 18(January).

Covshoff, S., Burgess, S. J., Kneřová, J. and Kümpers, B. M. C. (2014) 'Getting the most out of natural variation in C4 photosynthesis', *Photosynthesis Research*, 119(1-2), pp. 157-167.

Crider, K. S., Yang, T. P., Berry, R. J. and Bailey, L. B. (2012) 'Folate and DNA Methylation: A Review of Molecular Mechanisms and the Evidence for Folate's Role', *Advances in Nutrition*, 3(1), pp. 21-38.

Dai, X., Tu, X., Du, B., Dong, P., Sun, S., Wang, X., Sun, J., Li, G., Lu, T., Zhong, S. and Li, P. (2022) 'Chromatin and regulatory differentiation between bundle sheath and mesophyll cells in maize', *Plant Journal*, 109(3), pp. 675-692.

Dickinson, P., Knerova, J., Szecowka, M., Stevenson, S., Burgess, S., Mulvey, H., Bagman, A.-M., Gaudinier, A., Brady, S. and Hibberd, J. (2020) 'A bipartite transcription factor module controlling expression in the bundle sheath of Arabidopsis thaliana', *Nature Plants*, 6(12), pp. 1468-1479.

Dickinson, P. J., Triesch, S., Schlüter, U., Weber, A. P. M. and Hibberd, J. M. (2023) 'A transcription factor module mediating C2 photosynthesis', *bioRxiv*, pp. 2023.09.05.556297-2023.09.05.556297.

Edwards, E. J. (2019) 'Evolutionary trajectories, accessibility and other metaphors: the case of C4 and CAM photosynthesis', *New Phytologist*, 223(4), pp. 1742-1755.

Emmerling, J. (2018) *Studies into the Regulation of C4 Photosynthesis – Towards Factors Controlling Bundle Sheath Expression and Kranz Anatomy Development.* PhD Thesis, Heinrich-Heine-Universität Düsseldorf.

Feldmann, K. A. and David Marks, M. (1987) 'Agrobacterium-mediated transformation of germinating seeds of Arabidopsis thaliana: A non-tissue culture approach', *MGG Molecular & General Genetics*, 208(1-2), pp. 1-9.

Gouy, M., Guindon, S. and Gascuel, O. (2010) 'SeaView Version 4: A Multiplatform Graphical User Interface for Sequence Alignment and Phylogenetic Tree Building', *Molecular Biology and Evolution*, 27(2), pp. 221-224.

Gowik, U., Burscheidt, J., Akyildiz, M., Schlue, U., Koczor, M., Streubel, M. and Westhoff, P. (2004) 'cis-Regulatory elements for mesophyll-specific gene expression in the C4 plant Flaveria trinervia, the promoter of the C4 phosphoenolpyruvate carboxylase gene', *The Plant cell*, 16(5), pp. 1077-1090.

Gowik, U., Schulze, S., Saladié, M., Rolland, V., Tanz, S. K., Westhoff, P. and Ludwig, M. (2017) 'A MEM1-like motif directs mesophyll cell-specific expression of the gene encoding the C4 carbonic anhydrase in Flaveria', *Journal of Experimental Botany*, 68(2), pp. 311-320.

Grandi, F. C., Modi, H., Kampman, L. and Corces, M. R. (2022) 'Chromatin accessibility profiling by ATAC-seq', *Nature Protocols*, 17(6), pp. 1518-1552.

Groth, M., Moissiard, G., Wirtz, M., Wang, H., Garcia-Salinas, C., Ramos-Parra, P. A., Bischof, S., Feng, S., Cokus, S. J., John, A., Smith, D. C., Zhai, J., Hale, C. J., Long, J. A., Hell, R., Díaz De La Garza, R. I. and Jacobsen, S. E. (2016) 'MTHFD1 controls DNA methylation in Arabidopsis', *Nature Communications 2016 7:1*, 7(1), pp. 1-13.

Guerreiro, R., Bonthala, V. S., Schlüter, U., Hoang, N. V., Triesch, S., Schranz, M. E., Weber, A. P. M. and Stich, B. (2023) 'A genomic panel for studying C3-C4 intermediate photosynthesis in the Brassiceae tribe', *Plant Cell and Environment*, 46(11), pp. 3611-3627.

Hahn, F. (2018) *Genome editing and establishment of efficient gene targeting approaches in Arabidopsis using the CRISPR/Cas9 system.* PhD Thesis, Heinrich-Heine-Universität Düsseldorf.

Holst, F., Bolger, A., Günther, C., Maß, J., Triesch, S., Kindel, F., Kiel, N., Saadat, N., Ebenhöh, O., Usadel, B., Schwacke, R., Bolger, M., Weber, A. P. M. and Denton, A. K. (2023) 'Helixer–de novo Prediction of primary eukaryotic gene models conbining deep learning and a Hidden Marcov Model', *bioRxiv*.

Hylton, C. M., Rawsthorne, S., Smith, A. M., Jones, D. A. and Woolhouse, H. W. (1988) 'Glycine decarboxylase is confined to the bundle-sheath cells of leaves of C3-C4 intermediate species', *Planta,* 175(4), pp. 452-459.

Jafari, M., Guan, Y., Wedge, D. C. and Ansari-Pour, N. (2021) 'Re-evaluating experimental validation in the Big Data Era: a conceptual argument', *Genome Biology,* 22(1), pp. 1-6.

Kindel, F., Triesch, S., Schlüter, U., Randarevitch, L. A., Reichel-Deland, V., Weber, A. P. M. and Denton, A. K. (2024) 'Predmoter-cross-species prediction of plant promoter and enhancer regions'.

Krueger, F. and Andrews, S. R. (2011) 'Bismark: A flexible aligner and methylation caller for Bisulfite-Seq applications', *Bioinformatics,* 27(11), pp. 1571-1572.

Lin, M.-Y. (2020) *Studies into the genetic architecture of C 3 -C 4 characteristics in Moricandia presented by.* PhD Thesis, Heinrich-Heine-Universität Düsseldorf.

Lu, Z., Hofmeister, B. T., Vollmers, C., Dubois, R. M. and Schmitz, R. J. (2016) 'Combining ATAC-seq with nuclei sorting for discovery of cis-regulatory regions in plant genomes', *Nucleic Acids Research,* 45(6), pp. 41-41.

Lyu, M.-J. A., Gowik, U., Kelly, S., Covshoff, S., Hibberd, J. M., Sage, R. F., Ludwig, M., Wong, G. K.-S., Westhoff, P. and Zhu, X.-G. (2021) 'The coordination of major events in C4 photosynthesis evolution in the genus Flaveria', *Scientific Reports,* 11(1), pp. 1-14.

Mallmann, J., Heckmann, D., Bräutigam, A., Lercher, M. J., Weber, A. P. M., Westhoff, P. and Gowik, U. (2014) 'The role of photorespiration during the evolution of C4 photosynthesis in the genus Flaveria', *eLife,* 2014(3), pp. 1-23.

Monson, R. K. and Rawsthorne, S. (2000) 'CO2 Assimilation in C3-C4 Intermediate Plants', pp. 533-550.

Morgan, C. L., Turner, S. R. and Rawsthorne, S. (1993) 'Coordination of the cell-specific distribution of the four subunits of glycine decarboxylase and of serine hydroxymethyltransferase in leaves of C3-C4 intermediate species from different genera', *Planta,* 190(4), pp. 468-473.

Ratti, E. (2015) 'Big Data Biology: Between Eliminative Inferences and Exploratory Experiments', *Philosophy of Science,* 82(2), pp. 198-218.

Rawsthorne, S. (1992) C3–C4 intermediate photosynthesis: linking physiology to gene expression. *The Plant Journal.*

Rawsthorne, S., Hylton, C. M., Smith, A. M. and Woolhouse, H. W. (1988a) 'Distribution of photorespiratory enzymes between bundle-sheath and mesophyll cells in leaves of the C3-C4 intermediate species Moricandia arvensis (L.) DC', *Planta,* 176(4), pp. 527-532.

Rawsthorne, S., Hylton, C. M., Smith, A. M. and Woolhouse, H. W. (1988b) 'Photorespiratory metabolism and immunogold localization of photorespiratory enzymes in leaves of C3 and C3-C4 intermediate species of Moricandia', *Planta,* 173(3), pp. 298-308.

Sage, R. F. (2004) 'The evolution of C4 photosynthesis', *New Phytologist,* 161(2), pp. 341-370.

Schlüter, U., Bouvier, J. W., Guerreiro, R., Malisic, M., Kontny, C., Westhoff, P., Stich, B. and Weber, A. P. M. (2023) 'Brassicaceae display variation in efficiency of photorespiratory carbon-recapturing mechanisms', *Journal of Experimental Botany,* 74(21), pp. 6631-6649.

Schlüter, U., Bräutigam, A., Gowik, U., Melzer, M., Christin, P. A., Kurz, S., Mettler-Altmann, T. and Weber, A. P. M. (2017) 'Photosynthesis in C3-C4 intermediate Moricandia species', *Journal of Experimental Botany,* 68(2), pp. 191-206.

Schulze, S., Westhoff, P. and Gowik, U. (2016) 'Glycine decarboxylase in C3, C4 and C3–C4 intermediate species', *Current Opinion in Plant Biology,* 31, pp. 29-35.

Schwacke, R., Ponce-Soto, G. Y., Krause, K., Bolger, A. M., Arsova, B., Hallab, A., Gruden, K., Stitt, M., Bolger, M. E. and Usadel, B. (2019) 'MapMan4: A Refined Protein Classification and Annotation Framework Applicable to Multi-Omics Data Analysis', *Molecular Plant,* 12(6), pp. 879-892.

Shi, X. and Bloom, A. (2021) 'Photorespiration: The Futile Cycle?', *Plants 2021, Vol. 10, Page 908,* 10(5), pp. 908-908.

Singh, P., Stevenson, S. R., Dickinson, P. J., Reyna-llorens, I., Tripathi, A., Reeves, G., Schreier, T. B. and Hibberd, J. M. (2023) 'C 4 gene induction during de-etiolation evolved through changes in cis to allow integration with ancestral C 3 gene regulatory networks', 9(13).

Stiehler, F., Steinborn, M., Scholz, S., Dey, D., Weber, A. P. M. and Denton, A. K. (2020) 'Helixer: Cross-species gene annotation of large eukaryotic genomes using deep learning', *Bioinformatics*.

Swift, J., Luginbuehl, L. H., Schreier, T. B., Donald, R. M., Lee, T. A., Nery, J. R., Ecker, J. R. and Hibberd, J. M. (2023) 'Single nuclei sequencing reveals C4 photosynthesis is based on rewiring of ancestral cell identity networks', *bioRxiv*, pp. 2023.10.26.562893-2023.10.26.562893.

Triesch, S., Denton, A. K., Bouvier, J. W., Buchmann, J. P., Reichel-Deland, V., Guerreiro, R. N. F. M., Busch, N., Schlüter, U., Stich, B., Kelly, S. and Weber, A. P. M. (2024) 'Transposable elements contribute to the establishment of the glycine shuttle in Brassicaceae species', *Plant Biology,* 26(2), pp. 270-281.

Wang, L., Czedik-Eysenberg, A., Mertz, R. A., Si, Y., Tohge, T., Nunes-Nesi, A., Arrivault, S., Dedow, L. K., Bryant, D. W., Zhou, W., Xu, J., Weissmann, S., Studer, A., Li, P., Zhang, C., LaRue, T., Shao, Y., Ding, Z., Sun, Q., Patel, R. V., Turgeon, R., Zhu, X., Provart, N. J., Mockler, T. C., Fernie, A. R., Stitt, M., Liu, P. and Brutnell, T. P. (2014) 'Comparative analyses of C4 and C3 photosynthesis in developing leaves of maize and rice', *Nature Biotechnology*.

10

Outlook

In the previous three manuscripts, an extensive analysis of the differential gene expression in C_3-C_4 intermediate Brassicaceae and the underlying genetic mechanisms was presented. Beginning with a holistic view on spatial gene expression patterns in the C_3-C_4 intermediate *Moricandia arvensis*, we observed that the expression domain of numerous genes shifted to the bundle sheath cells (BSC). This activation of the BSC is a common mechanism in the evolution of C_4 photosynthesis (Sage *et al.*, 2012; Lyu *et al.*, 2021) and is at the basis of the specialized ultrastructure and function of the BSC. We were not yet able to find the genetic regulators that determine the developmental differences between the analyzed C_3 and C_3-C_4 species. Neither could we observe differential expression of *trans*-factors in our species that would point to master regulators of anatomy. This also includes the complex gene regulatory networks that connect circadian rhythms with light perception and the coordinated expression of photosynthetic genes (Singh *et al.*, 2023). Further single-cell studies with higher read counts and the comparison to more closely related C_3 *Moricandia* species have the chance to mitigate this.

We reported a shift of several photorespiratory genes to the BSC, which is in agreement with previous studies and seems to be a common mechanism across C_3-C_4 intermediate species from different taxa (Sage, 2004; Heckmann *et al.*, 2013). This shift primarily affects the gene encoding the P-protein of the glycine decarboxylase complex, and, accompanied with anatomical adjustments, the shift of the P-protein might be sufficient to install a basic C_3-C_4 intermediate metabolism (Hylton *et al.*, 1988; Sage *et al.*, 2014; Lundgren, 2020; Oono *et al.*, 2022). Initial attempts to restrict the *Arabidopsis thaliana* P-protein expression to the BSC using genome engineering however did not lead to a decreased carbon compensation point as observed in C_3-C_4 intermediate species (Hahn, 2018). This highlights the significance of additional adaptations, such as adequate energy balancing and the back-shuttle of photorespiratory nitrogen to the MC (Mallmann *et al.*, 2014; Bräutigam *et al.*, 2016; Sedelnikova *et al.*, 2018).

The expression domain shift of genes to the BSC underlies the control of *cis*-regulatory elements (CREs), mostly residing in the upstream region of genes. It is widely agreed that the recruitment of genes into the gene regulatory networks controlling C_4 photosynthesis is largely driven by co-option of CREs (Hibberd *et al.*, 2010; Cao *et al.*, 2016; Lyu *et al.*, 2021). Co-option of CREs is an efficient evolutionary strategy as it can reuse existing genetic modules for an easy rewiring of modular gene regulatory networks. Following this strategy, genes can be added or removed from gene regulatory networks, or multiple networks can be joined with ease. Previous gene or even genome duplications can prevent existing gene functions from being disrupted (Monson, 2012; Huang *et al.*, 2021; Huang *et al.*, 2023; Lambret-Frotte *et al.*, 2024).

The co-option of CREs can be mediated by transposable elements (TEs). At the level of individual loci, we discovered a TE insertion upstream of the *GLDP1* gene, encoding the P-protein of the glycine decarboxylase complex, selectively in C_3-C_4 intermediate genomes. In doing so, we found a remarkably convergent evolutionary event in the polyphyletic Brassicaceae clade with C_3-C_4 intermediacy. First, this finding highlights the role of high-quality genome assemblies for functional and quantitative genetics studies. The panel of Brassicaceae genomes assembled and analyzed in this study (Guerreiro *et al.*, 2023) will further help to unravel genetic mechanisms contributing to C_3-C_4 intermediate traits in Brassicaceae. Second, the TE insertion detected in this study underlines the role of TEs in differential gene expression, especially in the evolution of complex traits (Emmerling, 2018). TE mutagenesis was recently proposed as a transgene-free alternative for plant breeding (Kirov, 2023). Towards this, novel machine learning tools such as *Helixer* (Stiehler *et al.*, 2020; Holst *et al.*, 2023) and

EDTA (Ou et al., 2019) will make the identification and correlation of genes and TEs in non-model species significantly easier.

We sought to describe the role of the *GLDP1* upstream TE insertion in differential spatial gene expression. We found a potentially complex regulation by genetic and epigenetic mechanisms where it is still beyond our knowledge if the TE in the *GLDP1* upstream region acts as a spacer or as a target site for DNA methylation influencing gene activity. It is also not known if the methylation density of the TE underlies control by environmental factors. It is well conceivable that stress conditions such as elevated temperature decrease the methylation density and lead to a more accessible functional promoter region or alter the chromatin structure in the *GLDP1* upstream regions. It is known that factors such as light intensity and time of day have a significant impact on differential partitioning of gene expression between the MC and BSC (Borba *et al.*, 2023; Singh *et al.*, 2023). Testing the *GLDP1*::*GUS* fusion lines developed in this study under various environmental conditions can shed light on a putative flexible mechanism of spatially differential *GLDP1* expression. In addition, *A. thaliana* mutant lines with deficiencies in *de novo* DNA methylation could be used as background lines for the *GUS* fusion experiments. Novel methods like single-cell methylation sequencing or chromatin conformation sequencing will shed light on the mechanisms behind BSC-specific *GLDP1* expression in C_3-C_4 intermediate species.

This study featured the *bona fide* first single-cell transcriptome study of a C_3-C_4 intermediate Brassicaceae. Further applications of single-cell sequencing to additional intermediate species will describe the natural variation this trait exhibits in greater detail. The species *Diplotaxis erucoides* will be especially interesting as it seems to show a different metabolic profile compared to other C_3-C_4 intermediate species (Schlüter *et* *al.*, 2023) and does not exhibit a TE insertion in the *GLDP1* promoter despite elevated expression of this gene in the BSC.

Ultimately, the insights gained into the genetic framework of C_3-C_4 intermediacy will serve to reconstruct the evolution of this specialized trait and help to understand the complex evolution of C_4 photosynthesis. These findings will eventually be integrated into crop breeding programs and, in doing so, help to create more photosynthetically efficient crops. This translation from curiosity-driven research in *A. thaliana* to trait improvement programs in crops is not straightforward and it is therefore crucial to identify the ortholog loci of regulatory regions with high resolution (Leijten *et al.*, 2018; Inzé *et al.*, 2022). The insights we gained into the genetics of C_3-C_4 intermediacy in the Brassicaceae family and the methods we developed, together with the democratization of genomic data access will pave the way to a second green revolution.

Literature

Borba, A. R., Reyna-Llorens, I., Dickinson, P. J., Steed, G., Gouveia, P., Górska, A. M., Gomes, C., Kromdijk, J., Webb, A. A. R., Saibo, N. J. M. and Hibberd, J. M. (2023) 'Compartmentation of photosynthesis gene expression in C4 maize depends on time of day', *Plant Physiology,* 193(4), pp. 2306-2320.

Bräutigam, A. and Gowik, U. (2016) 'Photorespiration connects C3 and C4 photosynthesis', *Journal of Experimental Botany,* 67(10), pp. 2953-2962.

Cao, C., Xu, J., Zheng, G. and Zhu, X. G. (2016) 'Evidence for the role of transposons in the recruitment of cis-regulatory motifs during the evolution of C4 photosynthesis', *BMC Genomics,* 17(1), pp. 1-11.

Emmerling, J. (2018) *Studies into the Regulation of C4 Photosynthesis – Towards Factors Controlling Bundle Sheath Expression and Kranz Anatomy Development.* PhD Thesis, Heinrich-Heine-Universität Düsseldorf.

Guerreiro, R., Bonthala, V. S., Schlüter, U., Hoang, N. V., Triesch, S., Schranz, M. E., Weber, A. P. M. and Stich, B. (2023) 'A genomic panel for studying C3-C4 intermediate photosynthesis in the Brassiceae tribe', *Plant Cell and Environment,* 46(11), pp. 3611-3627.

Hahn, F. (2018) *Genome editing and establishment of efficient gene targeting approaches in Arabidopsis using the CRISPR/Cas9 system.* PhD Thesis, Heinrich-Heine-Universität Düsseldorf.

Heckmann, D., Schulze, S., Denton, A., Gowik, U., Westhoff, P., Weber, A. P. M. and Lercher, M. J. (2013) 'Predicting C4 photosynthesis evolution: Modular, individually adaptive steps on a mount fuji fitness landscape', *Cell,* 153(7), pp. 1579-1579.

Hibberd, J. M. and Covshoff, S. (2010) 'The regulation of gene expression required for C4 photosynthesis', *Annual Review of Plant Biology,* 61, pp. 181-207.

Holst, F., Bolger, A., Günther, C., Maß, J., Triesch, S., Kindel, F., Kiel, N., Saadat, N., Ebenhöh, O., Usadel, B., Schwacke, R., Bolger, M., Weber, A. P. M. and Denton, A. K. (2023) 'Helixer–de novo Prediction of primary eukaryotic gene models conbining deep learning and a Hidden Marcov Model', *bioRxiv*.

Huang, C. F., Liu, W. Y., Lu, M. Y. J., Chen, Y. H., Ku, M. S. B. and Li, W. H. (2021) 'Whole-Genome Duplication Facilitated the Evolution of C4 Photosynthesis in Gynandropsis gynandra', *Molecular Biology and Evolution,* 38(11), pp. 4715-4731.

Huang, C. F., Liu, W. Y., Yu, C. P., Wu, S. H., Ku, M. S. B. and Li, W. H. (2023) 'C4 leaf development and evolution', *Current Opinion in Plant Biology,* 76(September), pp. 102454-102454.

Hylton, C. M., Rawsthorne, S., Smith, A. M., Jones, D. A. and Woolhouse, H. W. (1988) 'Glycine decarboxylase is confined to the bundle-sheath cells of leaves of C3-C4 intermediate species', *Planta,* 175(4), pp. 452-459.

Inzé, D. and Nelissen, H. (2022) 'The translatability of genetic networks from model to crop species: lessons from the past and perspectives for the future', *New Phytologist,* 236(1), pp. 43-48.

Kirov, I. (2023) 'Toward Transgene-Free Transposon-Mediated Biological Mutagenesis for Plant Breeding', *International Journal of Molecular Sciences,* 24(23).

Lambret-Frotte, J., Smith, G. and Langdale, J. A. (2024) 'GOLDEN2-like1 is sufficient but not necessary for chloroplast biogenesis in mesophyll cells of C4 grasses', *The Plant Journal,* 117(2), pp. 416-431.

Leijten, W., Koes, R., Roobeek, I. and Frugis, G. (2018) 'Translating Flowering Time from Arabidopsis thaliana to Brassicaceae and Asteraceae Crop Species', *Plants 2018, Vol. 7, Page 111,* 7(4), pp. 111-111.

Lundgren, M. R. (2020) 'C2 photosynthesis: a promising route towards crop improvement?', *New Phytologist,* 228(6), pp. 1734-1740.

Lyu, M.-J. A., Gowik, U., Kelly, S., Covshoff, S., Hibberd, J. M., Sage, R. F., Ludwig, M., Wong, G. K.-S., Westhoff, P. and Zhu, X.-G. (2021) 'The coordination of major events in C4 photosynthesis evolution in the genus Flaveria', *Scientific Reports,* 11(1), pp. 1-14.

Mallmann, J., Heckmann, D., Bräutigam, A., Lercher, M. J., Weber, A. P. M., Westhoff, P. and Gowik, U. (2014) 'The role of photorespiration during the evolution of C4 photosynthesis in the genus Flaveria', *eLife,* 2014(3), pp. 1-23.

Monson, R. (2012) 'Gene Duplication, Neofunctionalization and the Evolution of C4 Photosynthesis', *International Journal of Plant Sciences,* 164(May 2003).

Oono, J., Hatakeyama, Y., Yabiku, T. and Ueno, O. (2022) 'Effects of growth temperature and nitrogen nutrition on expression of C3–C4 intermediate traits in Chenopodium album', *Journal of Plant Research*.

Ou, S., Su, W., Liao, Y., Chougule, K., Agda, J. R. A., Hellinga, A. J., Lugo, C. S. B., Elliott, T. A., Ware, D., Peterson, T., Jiang, N., Hirsch, C. N. and Hufford, M. B. (2019) 'Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline', *Genome Biology,* 20(1).

Sage, R. F. (2004) 'The evolution of C4 photosynthesis', *New Phytologist,* 161(2), pp. 341-370.

Sage, R. F., Khoshravesh, R. and Sage, T. L. (2014) From proto-Kranz to C4 Kranz: Building the bridge to C 4 photosynthesis. *Journal of Experimental Botany.*

Sage, R. F., Sage, T. L. and Kocacinar, F. (2012) 'Photorespiration and the evolution of C4 photosynthesis', *Annual Review of Plant Biology,* 63, pp. 19-47.

Schlüter, U., Bouvier, J. W., Guerreiro, R., Malisic, M., Kontny, C., Westhoff, P., Stich, B. and Weber, A. P. M. (2023) 'Brassicaceae display variation in efficiency of photorespiratory carbon-recapturing mechanisms', *Journal of Experimental Botany,* 74(21), pp. 6631-6649.

Sedelnikova, O. V., Hughes, T. E. and Langdale, J. A. (2018) 'Understanding the genetic basis of C4 Kranz anatomy with a view to engineering C3 crops', *Annual Review of Genetics,* 52, pp. 249-270.

Singh, P., Stevenson, S. R., Dickinson, P. J., Reyna-llorens, I., Tripathi, A., Reeves, G., Schreier, T. B. and Hibberd, J. M. (2023) 'C 4 gene induction during de-etiolation evolved through changes in cis to allow integration with ancestral C3 gene regulatory networks', 9(13).

Stiehler, F., Steinborn, M., Scholz, S., Dey, D., Weber, A. P. M. and Denton, A. K. (2020) 'Helixer: Cross-species gene annotation of large eukaryotic genomes using deep learning', *Bioinformatics*.

11

Journal Version of Manuscript II

plant biology

Plant Biology ISSN 1435-8603

RESEARCH ARTICLE

Transposable elements contribute to the establishment of the glycine shuttle in Brassicaceae species

S. Triesch[1,2], A. K. Denton[1,2], J. W. Bouvier[3], J. P. Buchmann[2,4], V. Reichel-Deland[1],
R. N. F. M. Guerreiro[5], N. Busch[1], U. Schlüter[1,2], B. Stich[2,5], S. Kelly[3] & A. P. M. Weber[1,2]

1 Institute for Plant Biochemistry, Heinrich Heine University Düsseldorf, Düsseldorf, Germany
2 Cluster of Excellence on Plant Sciences (CEPLAS), Düsseldorf, Germany
3 Department of Biology, University of Oxford, Oxford, UK
4 Institute for Biological Data Sciences, Heinrich Heine University Düsseldorf, Düsseldorf, Germany
5 Institute for Quantitative Genetics and Genomics of Plants, Heinrich Heine University Düsseldorf, Düsseldorf, Germany

Keywords

Brassicaceae; C_2 photosynthesis; C_3-C_4 intermediate photosynthesis; glycine decarboxylase; glycine shuttle; transposon.

Correspondence

A. P. M. Weber, Institute for Plant Biochemistry, Heinrich Heine University Düsseldorf, Düsseldorf, Germany.
E-mail: andreas.weber@uni-duesseldorf.de

All supplemental tables can be found under: https://git.nfdi4plants.org/hhu-plant-biochemistry/triesch2023_brassicaceae_transposons.

Editor

H.-H. Kunz

Received: 8 September 2023;
Accepted: 15 November 2023

doi:10.1111/plb.13601

ABSTRACT

- C_3-C_4 intermediate photosynthesis has evolved at least five times convergently in the Brassicaceae, despite this family lacking *bona fide* C_4 species. The establishment of this carbon concentrating mechanism is known to require a complex suite of ultrastructural modifications, as well as changes in spatial expression patterns, which are both thought to be underpinned by a reconfiguration of existing gene-regulatory networks. However, to date, the mechanisms which underpin the reconfiguration of these gene networks are largely unknown.
- In this study, we used a pan-genomic association approach to identify genomic features that could confer differential gene expression towards the C_3-C_4 intermediate state by analysing eight C_3 species and seven C_3-C_4 species from five independent origins in the Brassicaceae.
- We found a strong correlation between transposable element (TE) insertions in *cis*-regulatory regions and C_3-C_4 intermediacy. Specifically, our study revealed 113 gene models in which the presence of a TE within a gene correlates with C_3-C_4 intermediate photosynthesis. In this set, genes involved in the photorespiratory glycine shuttle are enriched, including the glycine decarboxylase P-protein whose expression domain undergoes a spatial shift during the transition to C_3-C_4 photosynthesis. When further interrogating this gene, we discovered independent TE insertions in its upstream region which we conclude to be responsible for causing the spatial shift in *GLDP1* gene expression.
- Our findings hint at a pivotal role of TEs in the evolution of C_3-C_4 intermediacy, especially in mediating differential spatial gene expression.

INTRODUCTION

C_4 photosynthesis has convergently evolved more than 60 times in flowering land plants (Sage *et al.* 2012). C_4 photosynthesis functions as a biochemical carbon concentrating mechanism that reduces the rate of photorespiration and thereby increases photosynthetic efficiency. Species that perform C_4 photosynthesis are mainly found in warm, dry and high-light environments in which leaf internal CO_2 levels are frequently low and, by extension, the oxygenation to carboxylation ratio of Rubisco is elevated (Sage *et al.* 2012; Betti *et al.* 2016). Although C_4 photosynthesis has evolved independently in multiple disparate plant lineages, the complexity of the required anatomical, biochemical, and developmental adaptations makes engineering C_4 photosynthesis a difficult undertaking.

Plants that exhibit C_3-C_4 intermediate phenotypes are promising research subjects to study the early steps towards C_4 photosynthesis (Kennedy & Laetsch 1974; Schlüter & Weber 2016; Bellasio & Farquhar 2019; Lundgren 2020). C_3-C_4 intermediate species exhibit specialized anatomical traits and they differ from C_4 species as they do not possess a fully integrated C_4 cycle. C_3-C_4 intermediate traits are characterized by a lowered CO_2 compensation point (CCP), chloroplast and mitochondria-rich bundle-sheath cells (BSC) and, in some cases, an increased vein density (Dengler *et al.* 1994; Christin *et al.* 2011; Schlüter *et al.* 2017). A further trait that is commonly shared between C_3-C_4 intermediate species from independent origins is the photorespiratory glycine shuttle, sometimes referred to as C_2 photosynthesis (reviewed in Schlüter & Weber (2016)). This shuttle relies on the BSC-specific decarboxylation of photorespiratory glycine, leading to an elevated CO_2 concentration around Rubisco. By extension, this increased partial pressure of CO_2 around the site of its fixation leads to a higher frequency of the Rubisco carboxylation reaction compared to oxygenation reactions, thereby suppressing photorespiration and resulting in decreased CCP (Kennedy & Laetsch 1974; Monson & Edwards 1984; Schlüter *et al.* 2017).

Changes in the spatial and temporal patterns of gene expression are crucial for the evolution of C_3-C_4 intermediate photosynthesis (Hibberd & Covshoff 2010; Reeves *et al.* 2017). Previously, it has been shown that the BSC-specific decarboxylation of glycine is caused by the differential localization of the glycine decarboxylase complex (GDC). In C_3-C_4 intermediate species from the genera *Moricandia*, *Flaveria* and *Panicum*, the P-protein of the GDC is only observed in BSC mitochondria, but not in mesophyll cell (MC) mitochondria (reviewed in Schulze *et al.* (2016)). This is a notable example of convergent evolution, as these species belong to the distant families Brassicaceae, Asteraceae and Poaceae. In these plants, loss of the GDC P-protein from the MC restricts glycine decarboxylation to the BSC in C_3-C_4 intermediate species (Rawsthorne *et al.* 1988; Morgan *et al.* 1993; Schulze *et al.* 2016). However, the exact mechanism by which this is achieved differs between different species. For instance, in C_3 *Flaveria*, the gene encoding the GDC P-protein (GLDP) is present in two differentially regulated copies, *GLDPA* and *GLDPB*. In C_3-C_4 intermediate *Flaveria* species, the ubiquitously expressed *GLDPB* is downregulated compared to C_3 *Flaveria* species, whereas the BSC-specific *GLDPA* is highly expressed (Schulze *et al.* 2013). In contrast, in C_3-C_4 intermediate *Moricandia*, the differential expression of *GLDP* is thought to be mediated by the loss of one gene copy and a change in regulation of the other copy. Specifically, in C_3-C_4 intermediate Brassiceae species, *GLDP2* is absent and *GLDP1* was reported to be differentially expressed by loss of a potential *cis*-element called M-Box. The M-Box element in the *Arabidopsis thaliana GLDP1* promoter confers a low-level expression in both MC and BSC and is absent from the upstream region of *GLDP1* in C_3-C_4 intermediate *Moricandia* species. A second *cis*-element, the V-Box, was shown to confer high levels of expression in the BSC and is present in all analysed Brassicaceae *GLDP1* promoter sequences to date (Adwy *et al.* 2015, 2019). Thus, there are multiple mechanisms through which *GLDP1* expression can be changed, from being ubiquitously expressed in the leaf, to being BSC-specific in C_3-C_4 plants.

Structural variation can originate from the activity of mobile genetic elements. In plants, transposable elements (TEs) comprise a large fraction of mobile genetic elements and contribute substantially to genome size variation (Lee & Kim 2014) and have substantial effects on the expression of genes (Hirsch & Springer 2017). TEs can be divided into two classes (Wicker *et al.* 2007) based on their transposition mechanisms: Class I transposons proliferate *via* a "copy-and-paste" mechanism involving an RNA intermediate, whereas Class II transposons transpose directly *via* a "cut-and-paste" mechanism. Due to their impact on structural variation, it has been frequently proposed that TEs can play a part in genome evolution and the evolution of novel genetic and phenotypic features (Wicker *et al.* 2007; Feschotte 2008; Buchmann *et al.* 2012; Qiu & Köhler 2020). Decades ago, Britten & Davidson (1971) put forward the idea that co-option of mobile sequences containing gene regulatory elements can connect genes to the same gene regulatory networks. The co-option of TEs for regulatory purposes is called "exaptation" (Brosius & Gould 1992). In the present day, with the vast amount of genomic data available, a deeper understanding of the role of transposable elements in genetic regulation allows linking genomic mechanisms with the evolution of complex traits.

TEs can rewire gene regulatory networks using different modes of action and influence the interplay of regulatory proteins (*trans*-elements) and the DNA sequences they are binding to (*cis*-elements). One such mode of action is the exaptation of a *cis*-regulatory element (CRE) from a separate gene (Fig. 1). If

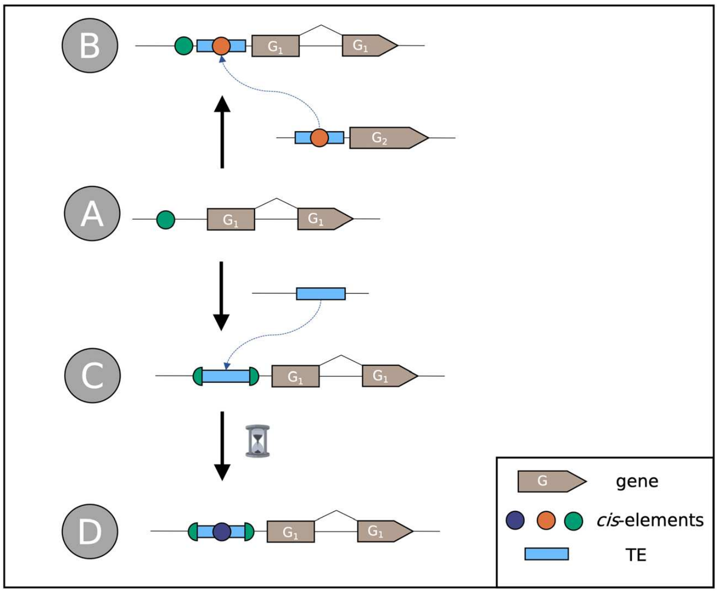


Fig. 1. Schematic illustration of gene regulation rewiring by TE exaptation. A: The hypothetical gene G1 is controlled by a *cis*-regulatory element (CRE, green dot). B: Gene G2 is regulated by a different CRE (orange dot) located within a TE (blue box). Upon transposition of the TE to the upstream region of G1, G1 might co-opt the function of the orange CRE, thus connecting G1 and G2 to the same gene regulatory network. C: TE transposition can also lead to destruction or suppression of the CRE. D: During TE decay, new CREs (blue dot) might occur through accumulation of point mutations.

the CRE inside a TE is copied from one gene and retained by the other gene, both genes become controlled by a mutual CRE and are thus connected by a shared gene regulatory network (Fig. 1B). In contrast to this scenario, it is also possible that TE integration into a CRE can suppress its function, either by interrupting the CRE sequence or altering the chromatin state of the respective CRE locus (Fig. 1C) (Feschotte 2008). A further possibility is the *de novo* generation of new CRE by point mutations in TEs (Fig. 1D). New CREs, *e.g.*, a 10-mer promoter element, can arise by random point mutations between 700,000 and 4.8 million years (Behrens & Vingron 2010).

Several examples for the role of TEs in rewiring gene regulatory networks in plants have been reported. In rice, the *mPing* DNA transposon was found preferentially in the 5′ region and was associated with the upregulation of stress response genes (Naito *et al.* 2009). In Brassicaceae, the evolution of heat tolerance was linked to the activity of *Copia* retrotransposons containing heat-shock factor binding elements (Pietzenuk *et al.* 2016). Furthermore, TEs were also found to be associated with endosperm development, *e.g.* the distribution of the PHEREs1 MADS-box transcription factor binding motifs by *Helitron* transposons in *A. thaliana* (Batista *et al.* 2019). The *Youren* miniature inverted-repeat TE (*MITE*) was shown to be transcribed in rice endosperm, putatively mediated by a NUCLEAR FACTOR Y binding motif in the vicinity of the 5′ terminal inverted repeat (TIR) of *Youren* (Nagata *et al.* 2022).

Previously, it has been shown that TEs play a significant role in the evolution of C_4 photosynthesis in maize. For instance, by analysing 40 C_4 gene orthologs between rice and maize for the presence of BSC-specific promoter motifs, Cao *et al.* (2016) identified over 1,000 promoter motifs that were differentially distributed between C_3 and C_4 orthologs, of which more than 60% were found to be associated with TEs and potentially co-opted by TE integration. These motifs may originate from non-photosynthetic genes and transposed to C_4 genes, which connected gene regulatory networks. The authors showed that TEs play a significant role in the evolution of C_4 photosynthesis in maize. However, the study of Cao *et al.* (2016) focused on evolutionary distant grasses, which makes it difficult to draw conclusions about the early evolutionary events towards C_4 photosynthesis.

In the present study, we test whether TE insertions are involved in decisive steps of the evolutionary establishment of C_3-C_4 intermediate photosynthesis. To do this, we focused on the Brassicaceae family which exhibits at least five independent origins of C_3-C_4 intermediate photosynthesis (Schlüter *et al.* 2022; Guerreiro *et al.* 2023) and contains multiple important and well-studied model plant species such as *A. thaliana*, *Arabis alpina* as well as relevant crop and vegetable plants such as *Brassica oleracea* (cabbage) and *Diplotaxis tenuifolia* (arugula).

We performed a pan-genomic association study to analyse the TE landscape of 15 Brassicaceae species. We tested for correlations between TE positions and the presence of C_3-C_4 intermediate traits. Specifically, we tested for correlations between the presence or absence of upstream co-occurring TEs with the CO_2 compensation point. In this unbiased approach, we aimed at finding genes that retained upstream TEs selectively only in C_3-C_4 intermediate plants. Based on the results of this analysis, we examined the upstream regions of relevant photorespiratory genes in closer detail to assess the potential role that TE insertions have played during establishment of C_3-C_4 photosynthesis traits. In doing so, we present evidence that the insertion of TEs in *cis*-regulatory regions of key genes is associated with the evolution of C_3-C_4 photosynthesis in the Brassicaceae.

MATERIAL AND METHODS

Genomes and carbon compensation points

The genomes of *Brassica gravinae* (Bg), *B. tournefortii* (Bt), *Carrichtera annua* (Ca), *Diplotaxis erucoides* (De), *D. tenuifolia* (Dt), *D. viminea* (Dv), *Hirschfeldia incana* (accessions HIR1 and HIR3), *Moricandia nitens* (Mn) and *M. suffruticosa* (Ms) were obtained from Guerreiro *et al.* (2023). The genome of *Arabis alpina* (Aa) was obtained from Jiao *et al.* (2017). The genome of *A. thaliana* (At) was obtained from Lamesch *et al.* (2012). The genome of *Moricandia arvensis* (Ma) and *M. moricandioides* (Mm) were obtained from Lin *et al.* (2021). The genome assembly for *Brassica oleracea* (Bo) was obtained from Parkin *et al.* (2014). The genome for *Gynandropsis gynandra* (Gg) was obtained from Hoang *et al.* (2022). A full list of species names and accession number and sources can be found in Table S1. Gas exchange data were obtained from Schlüter *et al.* (2022). The phylogenetic tree of all studied species was obtained from Guerreiro *et al.* (2023).

Gene annotation

Consistent structural gene annotations were generated for each species using *Helixer* (Holst *et al.* 2023) with the hybrid convolutional and bidirectional long-short term memory model, HybridModel, specifically the trained instance of land_plant_v0.3_m_0100 with default parameters.

Annotation of transposable elements

The TEs were *de novo* annotated using *EDTA* 1.9.9 (Ou *et al.* 2019) using the -anno 1 and -sensitive 1 flags. For the calculation of genomic composition (Figs 2 and 3), intact and fragmented TEs were used. To reduce the influence of false-positive hits, the pan-genomic gene-TE association study was performed for intact TEs only. The long terminal repeats (LTR) insertion time was calculated using

$$t_{\text{insertion}} = \frac{1 - \text{LTRidentity}}{2 \text{ x } \mu}$$

assuming a neutral mutation rate of $\mu = 1.4 \text{ x } 10^{-8}$ substitutions per site per year (Cai *et al.* 2018). The LTR identity was calculated as fraction of conserved base pairs of the aligned LTRs from the identified LTR elements:

$$\text{LTR identity} = \frac{\text{Number of conserved bp}}{\text{Number of total bp}}$$

Analysis of differential transposable element insertion

All downstream analyses were performed using *Python* 3.6 including *pandas* 1.2.4, *numpy* 1.20.1, *matplotlib* 3.4.1, *scikit-learn* 0.24.1, *scipy* 1.6.2 and *statsmodels* 0.12.2. All raw data and analyses are available in an Annotated Research Context (ARC)

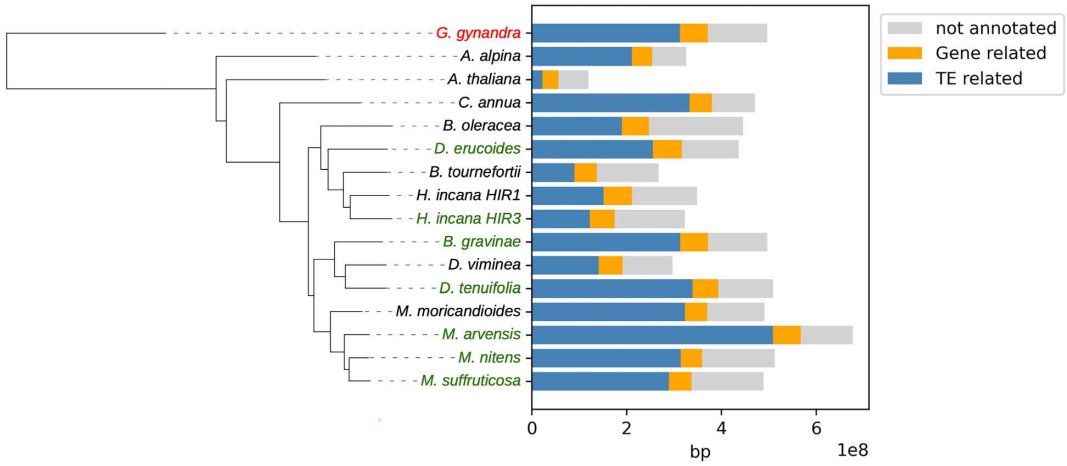


Fig. 2. Phylogeny and genomic composition of 15 selected Brassicaceae species and the Cleomaceae outgroup. C_3-C_4 intermediate species are highlighted in green, the C_4 outgroup *Gynandropsis gynandra* is highlighted in red. TE-related nucleotides are defined as spanning intact and fragmented transposon.

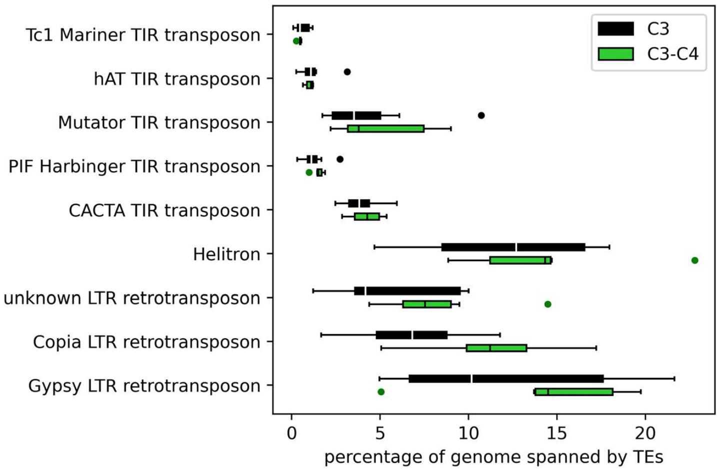


Fig. 3. Boxplot indicating the percentage of the genome comprised by each class of intact and fragmented TEs in eight C_3 and six C_3-C_4 intermediate species. The y-axis shows the TE classes, the x-axis indicates the fraction of the genome made up by the respective TE class. Black boxes depict C_3 species and green boxes depict C_3-C_4 intermediate species.

format under https://git.nfdi4plants.org/setri100/triesch2023_brassicaceae_transposons. A schematic workflow can be found in Supplementary Figure S1. The annotation files for genes and intact TEs were compared for each species. TEs were considered co-occurring with genes if their position matched one of the five cases described in Fig. 5. *CoGe SynMap* (https://genomevolution.org/coge/SynMap.pl) was used to identify orthologs and paralogs between the set of species. Each syntenic gene model was functionally annotated using *Mercator* 4.0 (Schwacke *et al.* 2019).

For each obtained syntelog, the effect of the presence or absence of an upstream TE on CCP was assessed using a phylogenetic implementation of the one-way ANOVA which accounts for the non-independence between species on the phylogenetic tree. For this purpose, phylogenetic ANOVAs were performed in the R environment using the *phylANOVA* function in the *phytools* 1.0.3 package (Revell 2012) using 1000 simulations and integrated post-hoc comparisons to evaluate differences between means.

Enrichment of *Mercator* bins for genes with correlating upstream TEs was calculated using Fisher's exact test. The identities of TEs in the *GLDP1* promoter were validated using the *CENSOR* webtool (Kohany *et al.* 2006).

RESULTS

The TE landscape of C_3 and C_3-C_4 Brassicaceae species

To screen for genomic features of potential relevance to the evolution of the C_3-C_4 photosynthesis trait, we conducted a pan-genomic association study of eight C_3 Brassicaceae species, seven C_3-C_4 intermediate Brassicaceae species from five independent origins, and one C_4 Cleomaceae as an outgroup species for tree building. The five independent origins of C_3-C_4 intermediate photosynthesis can be found in the *Moricandia arvensis*, *M. nitens*, and *M. suffruticosa* monophylum, as well as in *Diplotaxis erucoides*, *D. tenuifolia*, *Brassica gravinae*, and *Hirschfeldia incana* HIR3 (Fig. 2) (Schlüter & Weber 2016; Schlüter *et al.* 2022; Guerreiro *et al.* 2023).

The species panel exhibits genome sizes ranging from 120 Mbp in *A. thaliana* to 677 Mbp in *M. arvensis*. We found no significant difference in genome size between species exhibiting either the C_3 or C_3-C_4 intermediate photosynthesis phenotype (Fig. 2; one-way ANOVA $P > 0.05$). We next *de novo* annotated TEs using the *EDTA* pipeline (Ou *et al.* 2019). Overall, the annotated fragmented and intact transposons made up between 18% of the genome in *A. thaliana* and 75% in *M. arvensis*. We observed differences in genome size and TE content also in closely related species, between *M. arvensis* and *M. moricandioides* and between *B. gravinae* and *D. viminea*. Furthermore, we observed that differences in genome size are mainly due to the different TE content.

Class I type retrotransposons represented the majority of identified TEs across both C_3 and C_3-C_4 species (Fig. 3). For instance, across all analysed genomes, between 60% and 68% of all annotated TEs were Class I retrotransposons. In contrast, the proportion of TE classes in the genomes varied greatly across species (Fig. 3, Table S2).

The TE Class II was dominated by TEs from the *Helitron* group, making up between 5% and 20% of the genome (Fig. 3). The percentage of the genome made up of TEs from the different classes varied between the photosynthesis types, with a significantly higher amount of TEs in C_3-C_4 genomes (two-way ANOVA, $P = 0.013$).

To analyse recent increases of TE activity and their potential roles in the evolution of C_3-C_4 intermediate photosynthesis, we determined the insertion times of long terminal repeat (LTR) transposons (Fig. 4, Table S3). LTR retriever, which is the LTR annotation tool of the *EDTA* pipeline, detected LTR transposons to a threshold for repeat identity of 91%. Assuming a neutral mutation rate of $\mu = 1.4 \times 10^{-8}$ substitutions per site per year (Cai *et al.* 2018), LTR insertion times could thus be dated to a maximum of 4 million years ago. In general, both C_3 and C_3-C_4 intermediate species revealed the same broad pattern of LTR bursts. Specifically, in both groups, there was an increased frequency for LTR-TEs younger than 2 million years. However, the increase was more pronounced for C_3-C_4 intermediate species, largely on account of the high number of young LTR-TEs in *M. arvensis*. Statistical analysis revealed a significant correlation between the age distribution of LTR-transposons and the photosynthesis phenotype (two-way ANOVA, $P = 0.033$).

Upstream TEs are prevalent in C_3 and C_3-C_4 intermediate genomes

To better understand whether the high abundance of TEs in C_3-C_4 species was global or associated with specific genes, we next analysed the differential co-occurrence of TEs with protein coding genes. Co-occurrent TEs were defined as follows (Fig. 5): (I) the TE starts or ends in a 3,000 bp window upstream of the gene (upstream), (II) the TE starts or ends in a 3,000 bp window downstream of the gene (downstream), (III) the TE is residing within an exon or intron of the gene (inside), (IV) the TE starts but only partially resides in the gene (start), or (V) the TE ends but only partially resides in the gene (end).

Genes with TEs within the gene model (III) and overlapping TEs (IV and V) might have broken coding sequences and may result from imprecise annotations. Across the selected 11 species, 55,148 TEs were identified to be co-occurring with a protein coding gene in at least one species, whereas 21,643 co-occurring TEs belonged to C_3 and 28,379 co-occurring TEs belonged to C_3-C_4 species. In both C_3 and C_3-C_4 intermediate species, over 50% of the TEs co-occurring with genes were

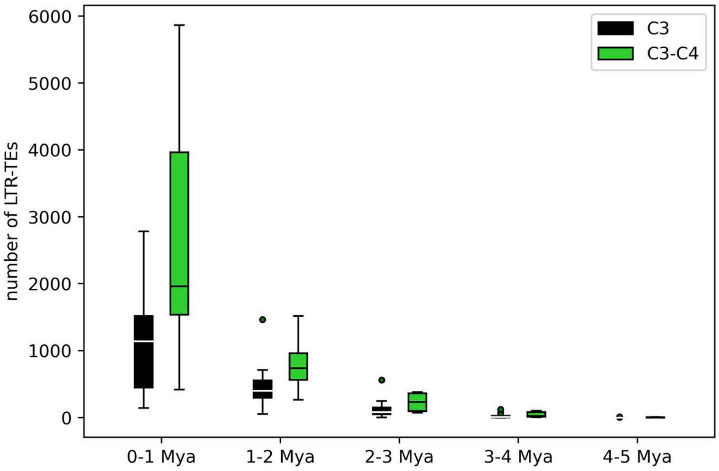


Fig. 4. Boxplot of LTR-TE insertion times for eight C_3 and six C_3-C_4 intermediate species. The x-axis shows the insertion time in bins of 1 million years before today (Mya). The y-axis depicts the number of identified LTR-TEs calculated to be inserted within this time frame. Calculation was performed using the LTR similarity of each LTR-TE and a neutral mutation rate of 1.4×10^{-8} substitutions per site per year. Black boxes represent C_3 species, green boxes represent C_3-C_4 species.

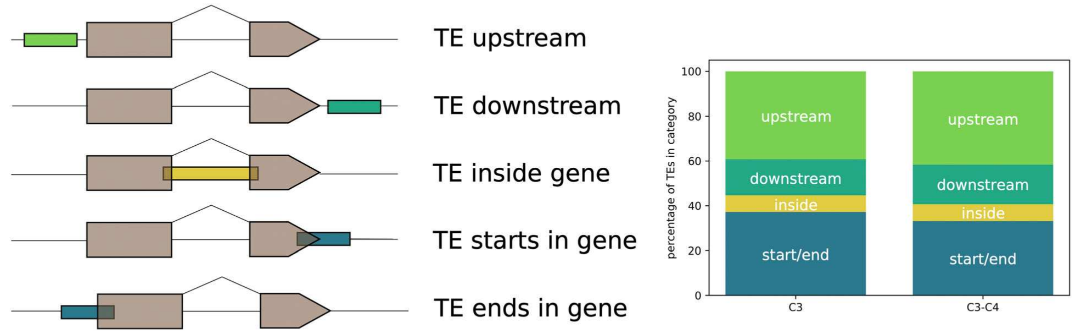


Fig. 5. Left panel: Different contexts of TEs co-occurring with genes. Right panel: Bar charts indicating the fractions of TE co-occurring with genes within five contexts: starting or ending in a gene (start/end), residing within a gene (inside) or residing within a 3000 bp window upstream or downstream the gene.

located up- or downstream of the gene (Fig. 5). Analysing potentially exaptated CREs, we focused on the up to 3000 bp 5′ region of the gene. To compare differential TE insertions between the analysed species, we obtained syntenic gene information for *CoGe SynMap*. For each of these syntenic gene models, one-way ANOVA was employed, correlating the presence or absence of a co-occurring upstream TE with the CCP of the respective species. After correcting the *P*-values for the phylogenetic bias, we identified 113 genes where the co-occurrence of one of the gene with an upstream TE correlated with the CCP ($P \leq 0.05$; Table 1, Table S4). Among the top ten genes (ranked by statistical confidence) were genes involved in photorespiration, such as the genes encoding the T- and P-subprotein of the glycine decarboxylase complex (Fig. 6A). Strikingly, the C_3-C_4 intermediate orthologs of these genes exhibited upstream TEs, whereas the C_3 orthologs lacked upstream TEs. Thus, during the evolution of C_3-C_4, there was a "gain" in upstream TEs in genes that function in photorespiration (Fig. 6A). In the subset of genes which exhibit an association between the presence of an upstream TE and the plant CCP, two photorespiratory genes occurred (*GLDP*, *GLDT*). To quantify putative enrichment of certain gene ontologies, each gene was functionally annotated with a *Mercator* bin. Statistical enrichment analysis using Fisher's exact test revealed that the *Mercator* bin "Photosynthesis.Photorespiration" ($P = 0.002907$) was enriched in the set of genes that co-occur with upstream transposons (Table 2). The occurrence of this *Mercator* bin was increased 38-fold over the background, which is higher than for any other analysed *Mercator* bin (Table 2).

Table 1. Selected subset of ten genes with upstream TEs with the lowest *P*-values for their association with the CCP.

gene name	AGI locus code	*P*-value
Glycine dehydrogenase component P-protein of glycine cleavage system	AT4G33010	0.001
Negative on TATA-less (NOT2)	AT5G59710	0.003
Regulatory protein FLZ of SnRK1 complex	AT5G49120	0.004
Pectate lyase	AT5G63180	0.005
MATE efflux family protein	AT2G38510	0.005
CYCLIN D-type regulatory protein	AT4G34160	0.005
Regulatory protein FLZ of SnRK1 complex	AT5G47060	0.005
Phosphocholine phosphatase (PS2/PECP1)	AT1G17710	0.007
PLATZ transcription factor family protein	AT3G50808	0.007
U-box domain-containing E3 ubiquitin ligase	AT4G25160	0.007

The *GLDP1* upstream region shows independent TE insertions in C_3-C_4 intermediate genomes

As *GLDP* was the gene model with the strongest association between the presence of upstream TEs and CCP, and it is known that the differential expression of *GLDP* contributes to the establishment of the photorespiratory glycine shuttle (Monson & Edwards 1984; Rawsthorne *et al.* 1988; Schulze *et al.* 2013), we selected this gene for further analysis. Several studies about the underlying regulatory genetics of *GLDP* expression have been conducted before (Adwy *et al.* 2015, 2019; Schulze *et al.* 2016; Dickinson *et al.* 2020). Only one *GLDP* gene copy is present in species from the Brassiceae tribe that contains all known C_3-C_4 intermediate species of the Brassicaceae (Schlüter *et al.* 2017). In contrast, the other two photorespiratory genes with correlating upstream TEs (Table 1, Fig. 6A) are found in higher copy numbers, which complicates a detailed genetic analysis.

We found three independent TE insertions in the promoter of C_3-C_4 intermediate *GLDP1* orthologs. In *Diplotaxis tenuifolia* a *Mutator* TE starts at 1970 bp upstream of the *GLDP1* start codon. In *Hirschfeldia incana* HIR3 a TE of the *Helitron* class is located around 2240 bp upstream. In orthologs from the monophyletic clade *Moricandia arvensis*, *M. nitens* and *M. suffruticosa* a *MITE* DNA transposon was detected, starting 1950 bp upstream of the *GLDP1* start codon. We calculated the minimum timespan since the *MITE* insertion by pairwise multiple sequence alignments of the *MITE* in the three *Moricandia GLDP1* promoters using the neutral mutation rate formula that was also employed for the calculation of LTR ages. We found that the *GLDP1* promoter *MITE* was at least 6.5 million years old.

All three independent TE insertions are located around 100 bp downstream of the M-Box promoter motif. This motif was previously hypothesized to confer MC expression (Adwy *et al.* 2015) since truncation of the motif from the *AtGLDP1* promoter shifted GUS activity from the whole leaf apex to the veins. Furthermore, the M-Box was reported to be lost in

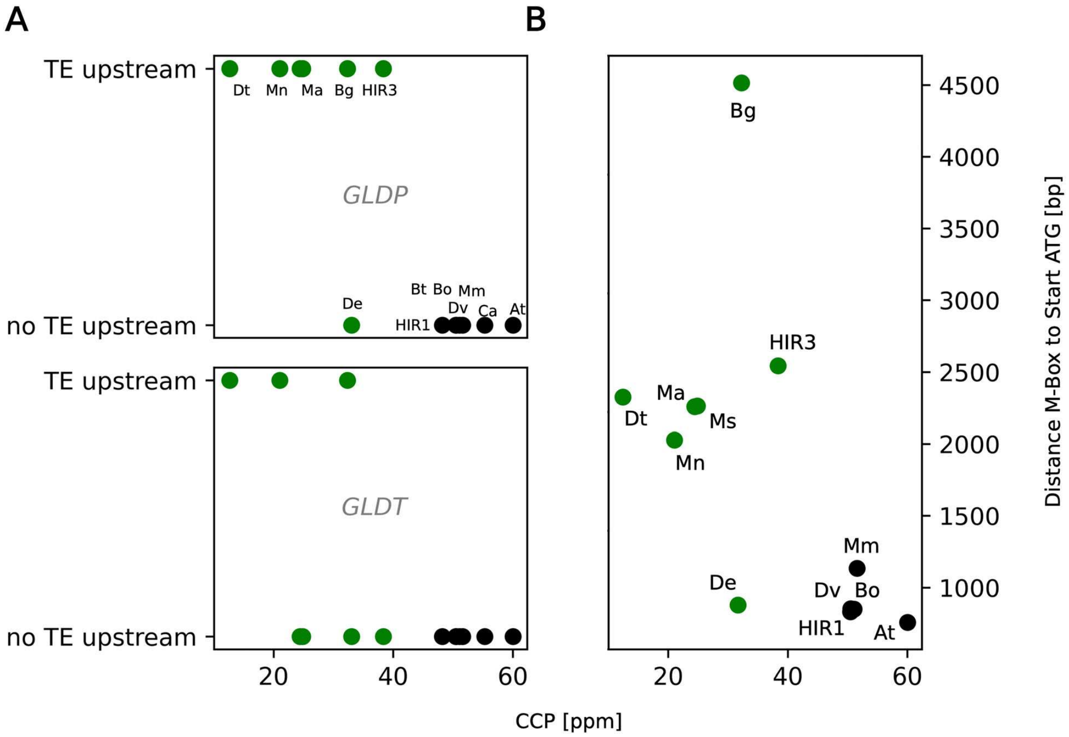


Fig. 6. A: Scatter plot for two photorespiratory genes with significant co-associated upstream TEs. The y-axis indicates the presence of an upstream TE (yes/no), the x-axis shows the carbon compensation point. Abbreviations: *GLDP/GLDT*: P/T-protein of the GLYCINE DECARBOXYLASE COMPLEX B: Scatter plot for the different architectures of the *GLDP1* promoter. The y-axis indicates the distance between the conserved M-Box sequence and the *GLDP1* start site. Each dot represents a species. C_3 species are shown in green, C_3-C_4 intermediate species are shown in black. Species name abbreviations: At: *Arabidopsis thaliana*, Bg: *Brassica gravinae*, Bo: *Brassica oleracea*, Bt: *Brassica tournefortii*, Ca: *Carrichtera annua*, De: *Diplotaxis erucoides*, Dt: *Diplotaxis tenuifolia*, Dv: *Diplotaxis viminea*, HIR1: *Hirschfeldia incana* HIR1, HIR3: *Hirschfeldia incana* HIR3, Ma: *Moricandia arvensis*, Mm: *Moricandia moricandioides*, Mn: *Moricandia nitens*, Ms: *Moricandia suffruticosa*.

Table 2. Results from two-sided Fisher's exact test for the enrichment of *Mercator* bins within the set of genes with significant upstream transposons.

Mercator bin	genes with $P > 0.05$	genes with $P < 0.05$	*P*-value	odds ratio
Photosynthesis.Photorespiration	3	2	0.002907	38.2
Multi-process regulation.SnRK1-kinase regulation	9	2	0.014932	12.7
Cell wall organization.cell wall proteins	32	3	0.022505	5.4
Solute transport.channels	45	3	0.050587	3.8

C_3-C_4 intermediate *Moricandia* species (Adwy *et al.* 2019). However, upon closer inspection, we found a highly conserved M-Box motif in all Brassicaceae genomes analysed here. Notably, the M-Box was shifted upstream due to the TE insertion in C_3-C_4 species, with the exception of *D. erucoides* (Figs 6B and 7, Table S5). In *Brassica gravinae*, the *EDTA* pipeline did not annotate an upstream transposon. However, we found a large insertion of unknown origin in the *B. gravinae GLDP1* promoter. This insertion is larger than the three reported TE cases but could be found in a similar position compared to the other *GLDP1* promoter insertions of TE origin (Fig. 7). In the *GLDP1* promoter of C_3-C_4 intermediate species *D. erucoides* no insertion could be found.

From five analysed C_3-C_4 *GLDP1* promoters, we found a large insertion behind the conserved M-Box in four cases (monophyletic C_3-C_4 intermediate *Moricandia* clade, *Diplotaxis tenuifolia*, *Brassica gravinae* and *Hirschfeldia incana* HIR3; Fig. 6B). Out of these four cases where the insertions occurred, we found evidence for the sequence being a TE in three cases (Fig. 7).

DISCUSSION

Individual TE insertions, not global TE patterns, are associated with C_3-C_4 intermediate photosynthesis

Evolution of new complex traits such as C_3-C_4 photosynthesis and C_4 photosynthesis requires the differential regulation of multiple genes. This includes differential gene

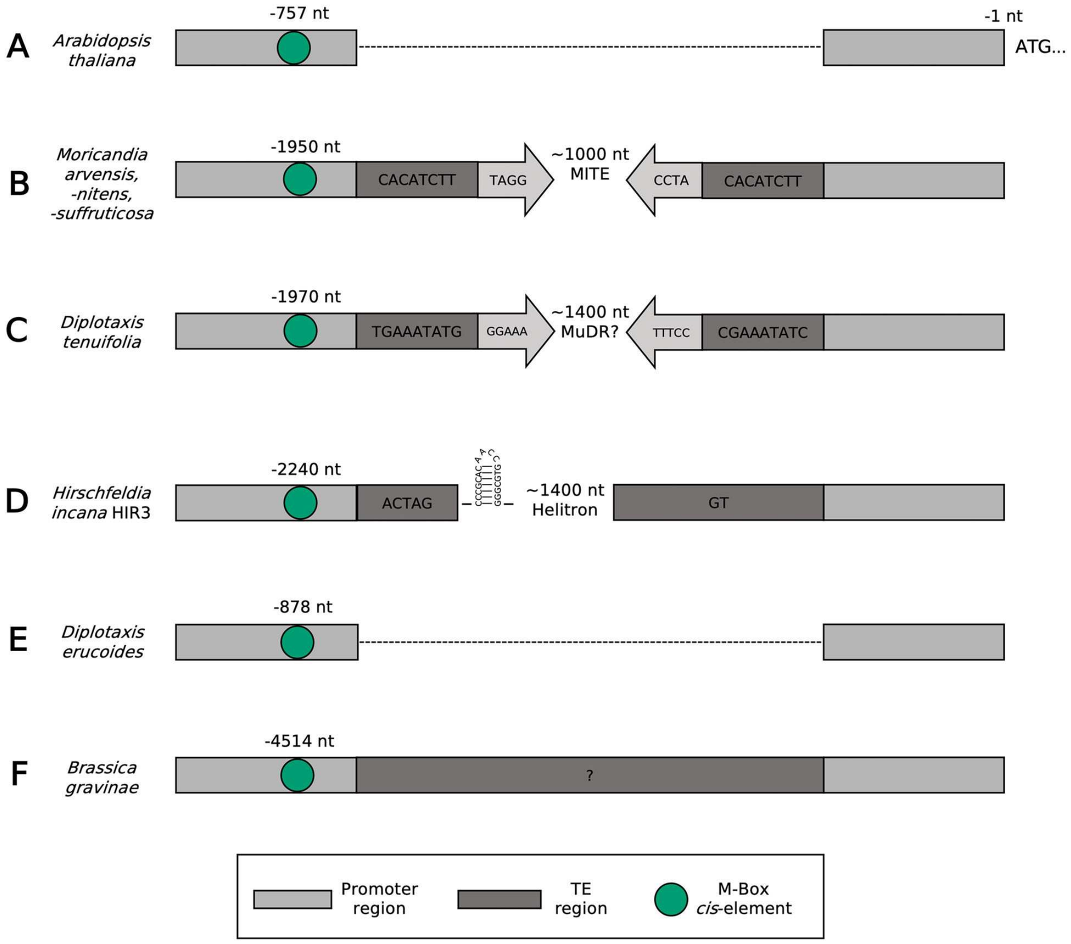


Fig. 7. Schematic representation of the *GLDP1* promoter region. "ATG…" depicts the start site of the *GLDP1* gene. Dark grey boxes represent characteristic TE sites such as target site duplications or the *Helitron* insertion sites. Grey arrows depict terminal inverted repeat motifs. The M-Box motif is highlighted as a green circle. In C_3 species such as *Arabidopsis thaliana* no TE is annotated in the promoter sequence, leading to a low spacing between the M-Box and the *GLDP1* start site (A). In the C_3-C_4 intermediate *Moricandia* species, a *MITE* TE begins around 1950 bp upstream of the *GLDP1* start codon (B). In *Diplotaxis tenuifolia*, a *Mutator* TE begins 1970 bp upstream (C). In *Hirschfeldia incana* HIR3 a *Helitron* with a highly conserved hairpin loop structure is inserted around 2240 bp upstream (D). Although being a C_3-C_4 intermediate species, the *Diplotaxis erucoides GLDP1* promoter did not have an insertion behind the M-Box. (E). In *Brassica gravinae* a large insertion of unknown origin could be found behind the M-Box region (F).

expression across both MSC and BSC tissue as well as the installation of light-responsiveness for genes of the core metabolism (reviewed in Hibberd & Covshoff (2010)). In many cases, the evolution of differential gene regulation takes place in promoter sequences, either by introduction or suppression of *cis*-elements.

A few *cis*-elements for MC specificity have been previously found, including the MEM1 motif from the *Flaveria trinervia* phosphoenolpyruvate carboxylase gene (Gowik *et al.* 2017) as well as the M-Box sequence in Brassicaceae (Adwy *et al.* 2015; Dickinson *et al.* 2020).

TEs have the potential to deliver or suppress *cis*-elements upon their insertion in a target promoter. TEs can generate antisense transcription, interrupt or generate heterochromatic regions, or serve as raw material for the *de novo* evolution of new *cis*-elements (reviewed in Feschotte (2008)). The role of TEs in the evolution of C_4 photosynthesis is only just starting to being uncovered. The present study comprises the first pan-genomic association analysis to assess the importance of TEs in the evolution of C_3-C_4 intermediacy. Specifically, to do this, we analysed the role of differential TE landscapes in 15 Brassicaceae species. First, we investigated whether genome size and TE content correlate with the presence of the C_3-C_4 photosynthesis phenotype. Across our species panel a variety of genome sizes is present (Fig. 2), but we could detect no correlation between genome size and the presence of the photosynthesis trait. However, it is possible that different levels of heterozygosity in the sequenced species may confound these results and genome size estimations have to be handled with care.

Within the Brassicaceae family species exhibiting C_3-C_4 intermediate traits can only be found in the Brassiceae tribe. Notably, species from this tribe seem to have undergone recent polyploidization events (Walden *et al.* 2020) and exhibit larger genome sizes than species from neighbouring tribes (Lysak *et al.* 2009).

Next, we analysed the proportion of TEs across individual genomes. Our estimation of TE proportions is consistent with previously analysed Brassicaceae genomes (Mirouze & Vitte 2014; Liu *et al.* 2020) and the *Gynandropsis gynandra* genome (Hoang *et al.* 2022). While genome size and TE content vary between species, we found a significant correlation between the photosynthesis phenotype and the proportion of the genome occupied by TEs in the respective species. Moreover, we found a recent burst in LTR-TE activity that is consistent with other studies (*e.g.*, Cai *et al.* 2018). The recent sharp increase in LTR-TE bursts in C_3-C_4 species comes mainly from *Moricandia arvensis* and might rather be due to high heterozygosity of LTR-containing genomic regions (Fig. 4). Although we found a significant correlation between LTR content and age with the C_3-C_4 intermediate phenotype, we cannot ultimately conclude that LTR transposon bursts contributed to the evolution of the C_3-C_4 intermediacy. Our LTR age analysis is limited to an LTR age of 4 million years. Given the estimated divergence time of 2–11 million years for C_3 and C_3-C_4 intermediate *Moricandia* species (Arias *et al.* 2014), our analysis of LTR insertion times will miss the contribution of older LTRs to the evolution of C_3-C_4 intermediate traits. Furthermore, based on sequence identity between the C_3-C_4 intermediate *Moricandia GLDP1* promoters, we estimate the age of the *MITE* in the *Moricandia GLDP1* promoter to be at least 6.5 million years. This also falls within the proposed divergence time C_3 and C_3-C_4 intermediate *Moricandia* species of 2–11 million years (Arias *et al.* 2014). Thus, changes in TE content occurred concomitant with the evolution of C_3-C_4 intermediate photosynthesis and occurred in genes whose expression is required to change for operation of a C_3-C_4 cycle.

In the descriptive whole-genome view, we observed correlations between TE content and age and the C_3-C_4 intermediate phenotype. Yet, however, there is an individual TE pattern even in closely related lines (Fig. 2). We therefore conclude that the role of TE activity may have an influence on C_3-C_4 evolution, but not necessarily *via* means of general TE activity (TE outbursts or TE purging) but rather *via* selective TE insertions to relevant genes or upstream regions. To analyse this, we employed a pan-genomic *de novo* transposon–gene association study, where we correlated the co-occurrence of TEs with genes to the presence of a C_3-C_4 intermediate phenotype.

In both C_3 and C_3-C_4 intermediate species, more than 50% of the analysed co-occurring TEs were upstream or downstream of the respective co-occurring gene or spanning the gene. This is biologically plausible, as TEs crossing gene borders may disturb gene function and intergenic regions can harbour transposable elements (Buchmann *et al.* 2012). Nevertheless, we found over 30% of the transposons crossing the borders of annotated genes. We assume that this was due to imprecise annotations by the TE identification pipeline.

Differential gene regulation mediated by variation in upstream regions was shown to be a driver of C_4 trait evolution in multiple, well documented cases (Wiludda *et al.* 2012; Adwy *et al.* 2015; Williams *et al.* 2015; Gowik *et al.* 2017). Our analysis revealed 113 genes with an upstream TE that correlates with the presence of a C_3-C_4 intermediate phenotype (Fig. 7; $P < 0.05$). Enrichment analysis of *Mercator* bins for this set of genes revealed an enrichment of the codes “Multi-process regulation.sucrose non-fermenting-related kinase (SnRK1) regulation” and “Photosynthesis.Photorespiration”. SnRK1 was shown to act as a central regulator of starvation metabolism that mediates energy homeostasis between organelles (Wurzinger *et al.* 2018). During nutrient starvation, SnRK1 subcomplexes were found to regulate the differential expression of over 600 target genes (Baena-González *et al.* 2007). Strikingly, ultrastructural adjustments and re-localization of the GDC P-protein to the BSC were demonstrated as a result of nitrogen starvation in the C_3-C_4 intermediate species *Chenopodium album* (Oono *et al.* 2022).

There is a clear bias of TE retention upstream of photorespiratory and SnRK1-regulatory genes in C_3-C_4 intermediate species, although with a small effect size (two out of five genes with $P < 0.05$ for “Photosynthesis. Photorespiration”; two out of 11 genes with $P < 0.05$ for “Multi-process regulation.SnRK1 regulation”; see Table 2).

We suggest that TE retention upstream of these genes has functional consequences, such as differential gene expression, putatively due to the co-option of new, or suppression of existing, *cis*-elements. Strikingly, the set of genes that are significantly enriched for the presence of TEs in the upstream region contains multiple genes involved in photorespiration, such as those encoding the T- and P- proteins of the glycine decarboxylase complex (GLDT/GLDP). The modification of photorespiration is an important step towards the establishment of the glycine shuttle. The enrichment of TE insertions upstream of photorespiratory genes in C_3-C_4 intermediates is a potential hint that TEs play a significant role in the introduction of the glycine shuttle.

The TE insertions in the *GLDP1* upstream region are highly convergent drivers of bundle-sheath cell specificity

The *GLDP* gene is a well-characterized example for differential gene expression at the early stages of C_3-C_4 evolution across multiple plant lineages (Schulze *et al.* 2013; Schlüter & Weber 2016). In the Brassiceae tribe, the *GLDP2* copy was lost (Schlüter *et al.* 2017). Additionally, *GLDP1* was reported to be differentially expressed between C_3 and C_3-C_4 intermediate *Moricandia* species (Hylton *et al.* 1988). In *A. thaliana*, GUS activity was restricted to the BSC by truncating the *GLDP1* promoter in the position of the M-Box, a promoter element *ca.* 800 bp upstream of the *AtGLDP1* gene start site. It was hypothesized that the M-Box confers MC expression, whereas expression in BSC is controlled by a MYC-MYB transcription factor binding module (Dickinson *et al.* 2023). Promoter-*GUS* fusions showed that the *GLDP1* promoter of the C_3 species *M. moricandioides* conferred *GUS* expression to both MC and BSC, whereas the *GLDP1* promoter of the C_3-C_4 intermediate species *M. arvensis* restricted *GUS* expression to the BSC (Adwy *et al.* 2019).

Adwy *et al.* (2019) explain the establishment of the glycine shuttle in *Moricandia* by the loss of the M-Box in C_3-C_4 intermediate *Moricandia* species. However, in contrast to this, we found the M-Box sequence in all our analysed *GLDP1* promoter variants, although this motif was shifted by over 1000 bp further upstream by the insertion of three independent TEs in the promoters in three independent evolutionary origins of C_3-C_4 intermediate photosynthesis, and by an insertion of unknown provenance in a fourth independent origin. This shift may have led to the M-Box being overlooked in previous studies.

Based on the findings of Adwy *et al.* (2019), we conclude that not the loss of the M-Box, but rather the upstream shift of the element by insertion of a TE has led to the differential tissue-specific expression of the *GLDP1* gene. The upstream shift of the M-Box was mediated by three independent TE insertions in lines with independent evolutionary origins of C_3-C_4 photosynthesis. This hints at a remarkable convergent evolutionary genetic mechanism in C_3-C_4 evolution. We suggest that the loss of *GLDP2* paved the way for neofunctionalization of the *GLDP1* copy in the Brassiceae tribe, the only Brassicaceae tribe containing C_3-C_4 intermediate species. This was mediated by the insertion of a TE in the promoter, suppressing the M-Box element and shifting *GLDP1* expression. It is questionable whether the TE insertion took place before or after the preconditioning of C_3-C_4 photosynthesis by anatomical adaptations, such as higher vein density and the distinct leaf anatomy. Hypothetically, limited expression of *GLDP1* in the MC may have been deleterious without further adaptations, which could have prevented the TE retention in the promoter. In *D. erucoides* we do not find a transposon in the *GLDP1* promoter region. The spacing of the M-Box to the *GLDP1* start codon is in the range of C_3 plants (Fig. 6B). However, *D. erucoides* shows C_3-C_4 intermediate phenotypes (Schlüter *et al.* 2017; Lundgren 2020). We assume that, being an independent evolutionary origin of C_3-C_4 intermediate photosynthesis, *D. erucoides* either shifted *GLDP1* expression to the BSC by different means or, alternatively, that there must be other additional regulators in the *GLDP1* promoter beyond our transposon-M-Box model. Contrasting the well-studied GDC activity and localization in *Moricandia* species, there are no data on the *D. erucoides* GDC biochemistry and genetics. Therefore, we cannot rule out that the glycine shuttle in *D. erucoides* is mediated by a different GDC regulation compared to the other C_3-C_4 intermediate species, such as the differential activity of the GDC T-, L-, or H- proteins.

By adopting a whole-genome view of TE density and gene–TE associations, our study highlights the potential importance of TE insertions in contributing to the convergent evolution of C_3-C_4 intermediacy. Differential *GLDP* expression is one of the most important innovations that occurs and facilitates the establishment of the glycine shuttle. The novel genetic mechanism of differential *GLDP1* regulation by a TE-mediated insertion causing an upstream shift of the M-Box must be verified in experimental work. The lack of efficient transformation protocols represents a significant impediment to functional genetics studies in non-model plants. Thus far, the successful transformation of any plant within our Brassicaceae species panel, apart from *A. thaliana*, has proven elusive, thereby precluding genomic engineering in C_3-C_4 intermediate Brassicaceae. The validation of the impact of TEs, for example on *GLDP1* expression *in planta*, hinges on the future accessibility of these species to genetic transformation. These experiments may necessitate the alteration of TE types or manipulating the positioning of CREs in upstream regions. For example, using a CRISPR-associated genomic engineering technique, TE insertions in upstream regions could be changed to different TE types, elongated, shortened or even relocated to downstream or intronic positions. Studying the influence of TEs on regulatory upstream regions *via* promoter–reporter studies can be conducted using transgenic *A. thaliana* lines. Nonetheless, it is imperative to consider that, due to their involvement in epigenetic regulation, particularly as hotspots for cytosine methylation, transgenic TEs may behave distinctly in transgenic *A. thaliana* when compared to their behaviour in their native host plant. Studying those genetic mechanisms of gene regulation in C_3-C_4 intermediate species will pave the way for a better understanding of the C_4 trait and facilitate genetic engineering efforts.

ACKNOWLEDGEMENTS

This work was funded by the Deutsche Forschungsgemeinschaft (German Research Foundation) under Germany's Excellence Strategy EXC-2048/1 under project ID 390686111, the Deutsche Forschungsgemeinschaft under Project ID 391465903/GRK 2466, the ERA-CAPS (European Research Network for Coordinating Action in Plant Sciences) project C4BREED under Project ID WE 2231/20-1 and the CRC (Collaborative Research Center) TRR341 under Project ID 456082119. JWB was funded by the BBSRC through BB/J014427/1. SK was funded by a Royal Society University Research Fellowship. Open Access funding enabled and organized by Projekt DEAL.

AUTHOR CONTRIBUTIONS

A.P.M.W., B.S. and U.S. designed and coordinated the project. S.T. designed and integrated all analyses. J.W.B and S.K. performed the phylogenetic correction of *P*-values. N.B. performed synteny analysis using *CoGe SynMap*. A.K.D. performed gene annotations using *Helixer*. A.K.D., R.N.F.M.G. and B.S. advised on statistical testing. All authors contributed to writing and accepted the manuscript.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Table S1. Overview over selected species with photosynthesis type and accession number or source.

Table S2. Number of nt spanned by intact and fragmented transposable elements per species analysed.

Table S3. Insertion times (age) of long terminal repeat transposons for each analysed species.

Table S4. Results of pan-genomic gene-transposon association study. Per gene, the absence (0) or presence (1) of a transposon within 3000 bp upstream of a gene is indicated for each analysed species. The AGI code represents the *A. thaliana* gene with the highest sequence homology.

Table S5. Distance of the M-Box to the *GLDP1* transcriptional start site for each analysed *GLDP1* ortholog upstream region.

Appendix S2. Supporting Information.

Figure S1. Flow chart depicting the computational workflow for the pan-genomic transposon-gene association study. File names highlighted in blue refer to scripts under https://git.nfdi4plants.org/hhu-plant-biochemistry/triesch2023_brassicaceae_transposons/-/tree/main/workflows. A: The *extensive de-novo TE annotator* (*EDTA*) software was used to annotate transposons in the selected genome sequences. EDTA distinguishes between intact and fragmented transposable elements (TEs). For the correlation of TEs and genes, only the intact TEs were used. Illustrated is one example TE (blue box) on a hypothetical contig at position 1000–2000 on the contig. *Helixer* was used to generate structural gene annotations. Depicted is one example gene (brown boxes) on a hypothetical contig at position 2500–5000 on the contig. B: Using a custom *python* script, the.gff3 files, containing the TE and gene annotations were compared and TE-gene associations as depicted in Fig. 5 were searched. In the example, the TE (blue box) resides up to 500 bp upstream of the example gene (brown box) and would thus be considered an upstream TE. C: For each genome, lists containing genes with TEs from the categories presented in Fig. 5 were created. The example from B would thus be appended to a list with genes that are associated with upstream TEs. *Mercator* was used to assign a functional annotation (*Mercator* bin) to all genes. Steps A–C were repeated for each genome. D: From the lists of genes with associated TEs per genome, a matrix was created where for each gene and species, the association of a gene with a TE was correlated with the carbon compensation point (CCP) of the species. These associations were tested using one-way ANOVA and resulting *P*-values were corrected for phylogenetic bias. Thus, a corrected *P*-value was assigned to each gene that indicated, whether there was a correlation of an associated TE with the CCP. E: From the *P*-values per gene, an arbitrary threshold of $P < 0.05$ was applied to divide the dataset. Fisher's test was used to quantify enrichment of *Mercator* bins within genes with $P < 0.05$.

REFERENCES

Adwy W., Laxa M., Peterhansel C. (2015) A simple mechanism for the establishment of C_2-specific gene expression in Brassicaceae. *The Plant Journal*, **84**, 1231–1238.

Adwy W., Schlüter U., Papenbrock J., Peterhansel C., Offermann S. (2019) Loss of the M-box from the glycine decarboxylase P-subunit promoter in C_2 *Moricandia* species. *Plant Gene*, **18**, 100176.

Arias T., Beilstein M.A., Tang M., McKain M.R., Pires J.C. (2014) Diversification times among Brassica (Brassicaceae) crops suggest hybrid formation after 20 million years of divergence. *American Journal of Botany*, **101**, 86–91.

Baena-González E., Rolland F., Thevelein J.M., Sheen J. (2007) A central integrator of transcription networks in plant stress and energy signalling. *Nature*, **448**, 938–942.

Batista R.A., Moreno-Romero J., Qiu Y., van Boven J., Santos-González J., Figueiredo D.D., Köhler C. (2019) The MADS-box transcription factor pheres1 controls imprinting in the endosperm by binding to domesticated transposons. *Elife*, **8**, e50541.

Behrens S., Vingron M. (2010) Studying the evolution of promoter sequences: a waiting time problem. *Journal of Computational Biology*, **17**, 1591–1606.

Bellasio C., Farquhar G.D. (2019) A leaf-level biochemical model simulating the introduction of C_2 and C_4 photosynthesis in C_3 rice: gains, losses and metabolite fluxes. *New Phytologist*, **223**, 150–166.

Betti M., Bauwe H., Busch F.A., Fernie A.R., Keech O., Levey M., Ort D.R., Parry M.A., Sage R., Timm S., Walker B., Weber A.P. (2016) Manipulating photorespiration to increase plant productivity: recent advances and perspectives for crop improvement. *Journal of Experimental Botany*, **67**, 2977–2988.

Britten R.J., Davidson E.H. (1971) Repetitive and non-repetitive DNA sequences and a speculation on the origins of evolutionary novelty. *The Quarterly Review of Biology*, **46**, 111–138.

Brosius J., Gould S.J. (1992) On 'genomenclature: a comprehensive (and respectful) taxonomy for pseudogenes and other 'junk DNA'. *Proceedings of the National Academy of Sciences of the United States of America*, **89**, 10706–10710.

Buchmann J.P., Matsumoto T., Stein N., Keller B., Wicker T. (2012) Inter-species sequence comparison of Brachypodium reveals how transposon activity corrodes genome colinearity. *The Plant Journal*, **71**, 550–563.

Cai X., Cui Y., Zhang L., Wu J., Liang J., Cheng L., Wang X., Cheng F. (2018) Hotspots of Independent and multiple rounds of LTR-retrotransposon bursts in Brassica species. *Horticultural Plant Journal*, **4**, 165–174.

Cao C., Xu J., Zheng G., Zhu X.G. (2016) Evidence for the role of transposons in the recruitment of cis-regulatory motifs during the evolution of C4 photosynthesis. *BMC Genomics*, **17**, 201.

Christin P.A., Sage T.L., Edwards E.J., Ogburn R.M., Khoshravesh R., Sage R.F. (2011) Complex evolutionary transitions and the significance of C_3–C_4 intermediate forms of photosynthesis in Molluginaceae. *Evolution*, **65**, 643–660.

Dengler N.G., Dengler R.E., Donnelly P.M., Hattersley P.W. (1994) Quantitative leaf anatomy of C3 and C4 grasses (Poaceae): bundle sheath and mesophyll surface area relationships. *Annals of Botany*, **73**, 241–255.

Dickinson P.J., Knerovà J., Szecowka M., Stevenson S.R., Burgess S.J., Mulvey H., Bagman A.M., Gaudinier A., Brady S.M., Hibberd J.M. (2020) A bipartite transcription factor module controlling expression in the bundle sheath of *Arabidopsis thaliana*. *Nature Plants*, **6**, 1468–1479.

Dickinson P.J., Triesch S., Schlüter U., Weber A.P., Hibberd J.M. (2023) A transcription factor module mediating C_2 photosynthesis bioRxiv, 2023-09.

Feschotte C. (2008) Transposable elements and the evolution of regulatory networks. *Nature Reviews Genetics*, **9**, 397–405.

Gowik U., Schulze S., Saladi'e M., Rolland V., Tanz S.K., Westhoff P., Ludwig M. (2017) A MEM1-like motif directs mesophyll cell-specific expression of the gene encoding the C_4 carbonic anhydrase in *Flaveria*. *Journal of Experimental Botany*, **68**, 311–320.

Guerreiro R., Bonthala V.S., Schlüter U., Hoang N.V., Triesch S., Schranz M.E., Weber A.P.M., Stich B. (2023) A genomic panel for studying C_3-C_4 intermediate photosynthesis in the Brassiceae tribe. *Plant, Cell & Environment*, **46**, 3611–3627. https://doi.org/10.1111/pce.14662

Hibberd J.M., Covshoff S. (2010) The regulation of gene expression required for C_4 photosynthesis. *Annual Review of Plant Biology*, **61**, 181–207. https://doi.org/10.1146/annurev-arplant-042809-112238

Hirsch C.D., Springer N.M. (2017) Transposable element influences on gene expression in plants. *Biochimica et Biophysica Acta Gene Regulatory Mechanisms*, **1860**, 157–165.

Hoang N.V., Sogbohossou E.O.D., Xiong W., Simpson C.J.C., Singh P., van den Bergh E., Zhu X.-G., Brautigam A., Weber A.P.M., van Haarst J.C., Schijlen E.G.W.M., Hendre P.S., Deynze A.V., Achigan-Dako E.G., Hibberd J.M., Schranz M.E. (2022) The genome of *Gynandropsis gynandra* provides insights into whole-genome duplications and the evolution of C_4 photosynthesis in Cleomaceae bioRxiv. 2022.07.09.499295.

Holst F., Bolger A., Günther C., Maß J., Triesch S., Kindel F., Kiel N., Saadat N., Ebenhöh O., Usadel B., Schwacke R., Bolger M., Weber A.P., Denton A.K. (2023) Helixer – de novo prediction of primary eukaryotic gene models conbining deep learning and a Hidden Marcov Model https://doi.org/10.1101/2023.02.06.527280.bioRxiv

Hylton C.M., Rawsthorne S., Smith A.M., Jones D.A., Woolhouse H.W. (1988) Glycine decarboxylase is confined to the bundle-sheath cells of leaves of C_3-C_4 intermediate species. *Planta*, **175**, 452–459.

Jiao W.B., Accinelli G.G., Hartwig B., Kiefer C., Baker D., Severing E., Willing E.M., Piednoel M., Woetzel S., Madrid-Herrero E., Huettel B., Hümann U., Reinhard R., Koch M.A., Swan D., Clavijo B., Coupland G., Schneeberger K. (2017) Improving and correcting the contiguity of long-read genome assemblies of three plant species using optical mapping and chromosome conformation capture data. *Genome Research*, **27**, 778–786.

Kennedy R.A., Laetsch W.M. (1974) Plant species intermediate for C_3, C_4 photosynthesis. *Science*, **184**, 1087–1089.

Kohany O., Gentles A.J., Hankus L., Jurka J. (2006) Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. *BMC Bioinformatics*, **7**, 474.

Lamesch P., Berardini T.Z., Li D., Swarbreck D., Wilks C., Sasidharan R., Muller R., Dreher K., Alexander D.L., Garcia-Hernandez M., Karthikeyan A.S., Lee C.H., Nelson W.D., Ploetz L., Singh S., Wensel A., Huala E. (2012) The Arabidopsis information resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Research*, **40**(D1), D1202–D1210.

Lee S.-I., Kim N.-S. (2014) Transposable elements and genome size variations in plants. *Genomics & Informatics*, **12**, 87–97.

Lin M.-Y., Koppers N., Denton A., Schlüter U., Weber A.P. (2021) Whole genome sequencing and assembly data of *Moricandia moricandioides* and *M. arvensis*. *Data in Brief*, **35**, 106922.

Liu Z., Fan M., Yue E.K., Li Y., Tao R.F., Xu H.M., Duan M.H., Xu J.H. (2020) Natural variation and evolutionary dynamics of transposable elements in *Brassica oleracea* based on next-generation sequencing data. *Horticulture Research*, **7**, 145.

Lundgren M.R. (2020) C_2 photosynthesis: a promising route towards crop improvement? *New Phytologist*, **228**, 1734–1740.

Lysak M.A., Koch M.A., Beaulieu J.M., Meister A., Leitch I.J. (2009) The dynamic ups and downs of genome size evolution in Brassicaceae. *Molecular Biology and Evolution*, **26**, 85–98.

Mirouze M., Vitte C. (2014) Transposable elements, a treasure trove to decipher epigenetic variation: insights from Arabidopsis and crop epigenomes. *Journal of Experimental Botany*, **65**, 2801–2812.

Monson R.K., Edwards G.E. (1984) C_3–C_4 intermediate photosynthesis in plants. *Bioscience*, **34**, 563–574.

Morgan C.L., Turner S.R., Rawsthorne S. (1993) Coordination of the cell-specific distribution of the four subunits of glycine decarboxylase and of serine hydroxymethyltransferase in leaves of C_3–C_4 intermediate species from different genera. *Planta*, **190**, 468–473.

Nagata H., Ono A., Tonosaki K., Kawakatsu T., Sato Y., Yano K., Kishima Y., Kinoshita T. (2022) Temporal changes in transcripts of miniature inverted-repeat transposable elements during rice endosperm development. *The Plant Journal*, **109**, 1035–1047.

Naito K., Zhang F., Tsukiyama T., Saito H., Hancock C.N., Richardson A.O., Okumoto Y., Tanisaka T., Wessler S.R. (2009) Unexpected consequences of a sudden and massive transposon amplification on rice gene expression. *Nature*, **461**, 1130–1134.

Oono J., Hatakeyama Y., Yabiku T., Ueno O. (2022) Effects of growth temperature and nitrogen nutrition on expression of C_3–C_4 intermediate traits in *Chenopodium album*. *Journal of Plant Research*, **135**, 15–27.

Ou S., Su W., Liao Y., Chougule K., Agda J.R., Hellinga A.J., Lugo C.S.B., Elliott T.A., Ware D., Peterson T., Jiang N., Hirsch C.N., Hufford M.B. (2019) Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biology*, **20**, 275.

Parkin I.A., Koh C., Tang H., Robinson S.J., Kagale S., Clarke W.E., Town C.D., Nixon J., Krishnakumar V., Bidwell S.L., Denoeud F., Belcram H., Links M.G., Just J., Clarke C., Bender T., Huebert T., Mason A.S., Chris Pires J., Barker G., Moore J., Walley P.G., Manoli S., Batley J., Edwards D., Nelson M.N., Wang X., Paterson A.H., King G., Bancroft I., Chalhoub B., Sharpe A.G. (2014) Transcriptome and methylome profiling reveals relics of genome dominance in the mesopolyploid *Brassica oleracea*. *Genome Biology*, **15**, R77.

Pietzenuk B., Markus C., Gaubert H., Bagwan N., Merotto A., Bucher E., Pecinka A. (2016) Recurrent evolution of heat-responsiveness in Brassicaceae COPIA elements. *Genome Biology*, **17**, 209.

Qiu Y., Köhler C. (2020) Mobility connects: transposable elements wire new transcriptional networks by transferring transcription factor binding motifs. *Biochemical Society Transactions*, **48**, 1005–1017.

Rawsthorne S., Hylton C.M., Smith A.M., Woolhouse H.W. (1988) Photorespiratory metabolism and immunogold localization of photorespiratory enzymes in leaves of C_3 and C_3–C_4 intermediate species of *Moricandia*. *Planta*, **173**, 298–308.

Reeves G., Grangè-Guermente M.J., Hibberd J.M. (2017) Regulatory gateways for cell-specific gene expression in C_4 leaves with Kranz anatomy. *Journal of Experimental Botany*, **68**, 107–116.

Revell L.J. (2012) phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, **3**, 217–223.

Sage R.F., Sage T.L., Kocacinar F. (2012) Photorespiration and the evolution of C_4 photosynthesis. *Annual Review of Plant Biology*, **63**, 19–47.

Schlüter U., Bouvier J.W., Guerreiro R., Malisic M., Kontny C., Westhoff P., Stich B., Weber A.P.M. (2022) Brassicaceae display diverse photorespiratory carbon recapturing mechanisms bioRxiv.

Schlüter U., Bräutigam A., Gowik U., Melzer M., Christin P.A., Kurz S., Mettler-Altmann T., Weber A.P. (2017) Photosynthesis in C_3–C_4 intermediate *Moricandia* species. *Journal of Experimental Botany*, **68**, 191–206.

Schlüter U., Weber A.P. (2016) The road to C_4 photosynthesis: evolution of a complex trait via intermediary states. *Plant and Cell Physiology*, **57**, 881–889.

Schulze S., Mallmann J., Burscheidt J., Koczor M., Streubel M., Bauwe H., Gowik U., Westhoff P. (2013) Evolution of C_4 photosynthesis in the genus *Flaveria*: establishment of a photorespiratory CO_2 pump. *The Plant Cell*, **25**, 2522–2535.

Schulze S., Westhoff P., Gowik U. (2016) Glycine decarboxylase in C_3, C_4 and C_3–C_4 intermediate species. *Current Opinion in Plant Biology*, **31**, 29–35.

Schwacke R., Ponce-Soto G.Y., Krause K., Bolger A.M., Arsova B., Hallab A., Gruden K., Stitt M., Bolger M.E., Usadel B. (2019) MapMan4: a refined protein classification and annotation framework applicable to multi-omics data analysis. *Molecular Plant*, **12**, 879–892.

Walden N., German D.A., Wolf E.M., Kiefer M., Rigault P., Huang X.C., Kiefer C., Schmickl R., Franzke A., Neuffer B., Mummenhoff K., Koch M.A. (2020) Nested whole-genome duplications coincide with diversification and high morphological disparity in Brassicaceae. *Nature Communications*, **11**, 3795.

Wicker T., Sabot F., Hua-Van A., Bennetzen J.L., Capy P., Chalhoub B., Flavell A., Leroy P., Morgante M., Panaud O., Paux E., SanMiguel P., Schulman A.H. (2007) A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics*, **8**, 973–982.

Williams B.P., Burgess S.J., Reyna-Llorens I., Knerova J., Aubry S., Stanley S., Hibberd J.M. (2015) An untranslated cis-element regulates the accumulation of multiple C_4 enzymes in *Gynandropsis gynandra* mesophyll cells. *The Plant Cell*, **28**, 454–465.

Wiludda C., Schulze S., Gowik U., Engelmann S., Koczor M., Streubel M., Bauwe H., Westhoff P. (2012) Regulation of the photorespiratory GLDPA gene in C_4 *Flaveria*: an intricate interplay of transcriptional and posttranscriptional processes. *The Plant Cell*, **24**, 137–151.

Wurzinger B., Nukarinen E., Nägele T., Weckwerth W., Teige M. (2018) The SnRK1 kinase as central mediator of energy signaling between different organelles. *Plant Physiology*, **176**, 1085–1094.