





## RESEARCH ARTICLE

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

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

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<sub>2</sub> photosynthesis; C<sub>3</sub>-C<sub>4</sub> 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](mailto:andreas.weber@uni-duesseldorf.de)

All supplemental tables can be found under: [https://git.nfdi4plants.org/hhu-plant-biochemistry/triesch2023\\_brassicaceae\\_transposons](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<sub>3</sub>-C<sub>4</sub> intermediate photosynthesis has evolved at least five times convergently in the Brassicaceae, despite this family lacking *bona fide* C<sub>4</sub> 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<sub>3</sub>-C<sub>4</sub> intermediate state by analysing eight C<sub>3</sub> species and seven C<sub>3</sub>-C<sub>4</sub> species from five independent origins in the Brassicaceae.
- We found a strong correlation between transposable element (TE) insertions in *cis*-regulatory regions and C<sub>3</sub>-C<sub>4</sub> intermediacy. Specifically, our study revealed 113 gene models in which the presence of a TE within a gene correlates with C<sub>3</sub>-C<sub>4</sub> 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<sub>3</sub>-C<sub>4</sub> 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<sub>3</sub>-C<sub>4</sub> intermediacy, especially in mediating differential spatial gene expression.

## INTRODUCTION

C<sub>4</sub> photosynthesis has convergently evolved more than 60 times in flowering land plants (Sage *et al.* 2012). C<sub>4</sub> photosynthesis functions as a biochemical carbon concentrating mechanism that reduces the rate of photorespiration and thereby increases photosynthetic efficiency. Species that perform C<sub>4</sub> photosynthesis are mainly found in warm, dry and high-light environments in which leaf internal CO<sub>2</sub> 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<sub>4</sub> photosynthesis has evolved independently in multiple disparate plant lineages, the complexity of the required anatomical, biochemical, and developmental adaptations makes engineering C<sub>4</sub> photosynthesis a difficult undertaking.

Plants that exhibit C<sub>3</sub>-C<sub>4</sub> intermediate phenotypes are promising research subjects to study the early steps towards C<sub>4</sub> photosynthesis (Kennedy & Laetsch 1974; Schlüter & Weber 2016; Bellasio & Farquhar 2019; Lundgren 2020). C<sub>3</sub>-C<sub>4</sub> intermediate species exhibit specialized anatomical traits and they differ from C<sub>4</sub> species as they do not possess a fully

integrated C<sub>4</sub> cycle. C<sub>3</sub>-C<sub>4</sub> intermediate traits are characterized by a lowered CO<sub>2</sub> 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<sub>3</sub>-C<sub>4</sub> intermediate species from independent origins is the photorespiratory glycine shuttle, sometimes referred to as C<sub>2</sub> photosynthesis (reviewed in Schlüter & Weber (2016)). This shuttle relies on the BSC-specific decarboxylation of photorespiratory glycine, leading to an elevated CO<sub>2</sub> concentration around Rubisco. By extension, this increased partial pressure of CO<sub>2</sub> 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).

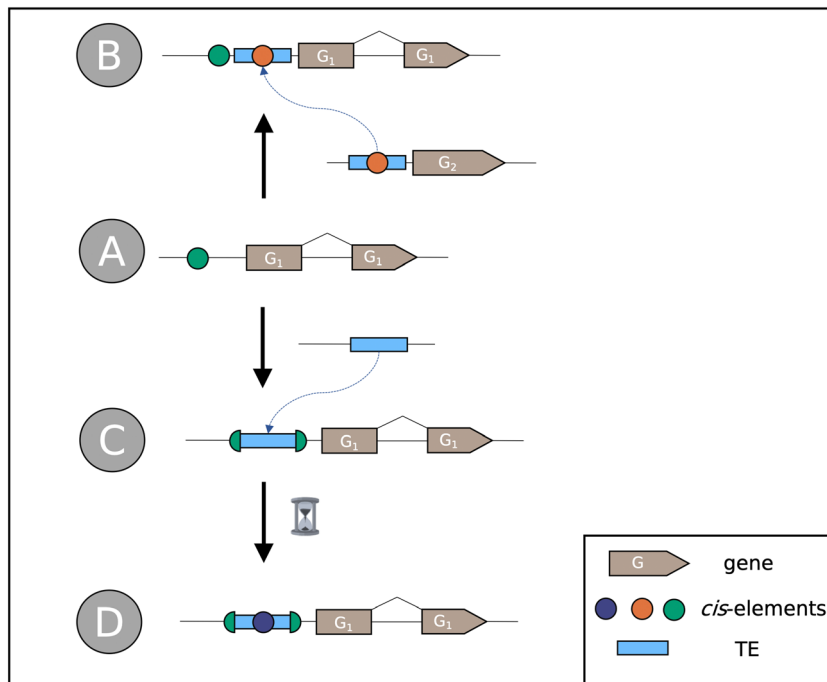
Changes in the spatial and temporal patterns of gene expression are crucial for the evolution of C<sub>3</sub>-C<sub>4</sub> intermediate photosynthesis (Hibberd & Covshoff 2010; Reeves *et al.* 2017). Previously, it has been shown that the BSC-specific decarboxylation of glycine is caused by the differential

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 Brassicaceae 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<sub>4</sub> photosynthesis in maize. For instance, by analysing 40 C<sub>4</sub> 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<sub>3</sub> and C<sub>4</sub> 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<sub>4</sub> genes, which connected gene regulatory networks. The authors showed that TEs play a significant role in the evolution of C<sub>4</sub> 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<sub>4</sub> photosynthesis.

In the present study, we test whether TE insertions are involved in decisive steps of the evolutionary establishment of C<sub>3</sub>-C<sub>4</sub> intermediate photosynthesis. To do this, we focused on the Brassicaceae family which exhibits at least five independent origins of C<sub>3</sub>-C<sub>4</sub> intermediate photosynthesis (Schlüter *et al.* 2022; Guerreiro *et al.* 2023) and contains multiple important and well-studied model plant species such as *A. thaliana*, *Arabidopsis thaliana* 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<sub>3</sub>-C<sub>4</sub> intermediate traits. Specifically, we tested for correlations between the presence or absence of upstream co-occurring TEs with the CO<sub>2</sub> compensation point. In this unbiased approach, we aimed at finding genes that retained upstream TEs selectively only in C<sub>3</sub>-C<sub>4</sub> 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<sub>3</sub>-C<sub>4</sub> 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<sub>3</sub>-C<sub>4</sub> 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 *Arabidopsis thaliana* (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

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

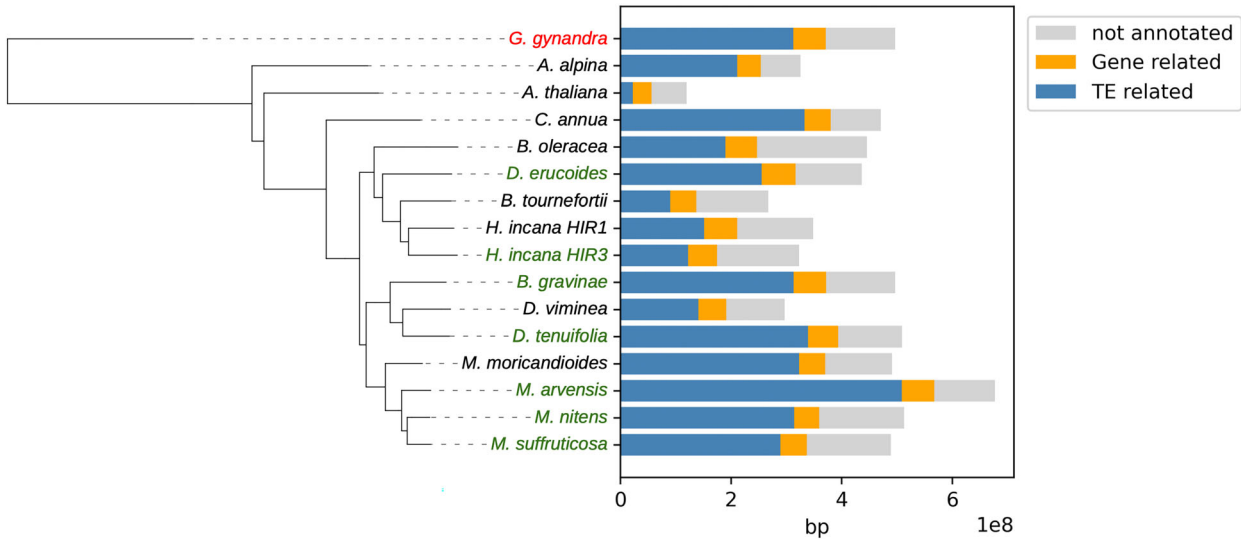
assuming a neutral mutation rate of  $\mu = 1.4 \times 10^{-8}$  substitutions per site per year (Cai *et al.* 2018). The LTR identity was calculated as fraction of conserved base pairs of the aligned LTRs from the identified LTR elements:

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

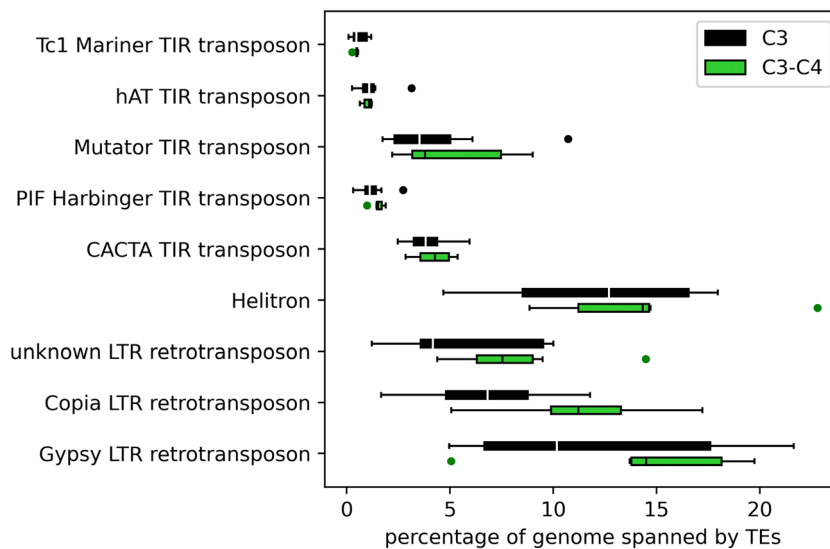
### Analysis of differential transposable element insertion

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)

Transposable elements contribute to the establishment of the glycine Triesch, Denton, Bouvier, Buchmann, Reichel-Deland, Guerreiro, Busch, Schlüter, Stich, Kelly & Weber



**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](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://genomeevolution.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<sub>3</sub>-C<sub>4</sub> intermediate Brassicaceae species from five independent origins, and one C<sub>4</sub> Cleomaceae as an outgroup species for tree building. The five independent origins of C<sub>3</sub>-C<sub>4</sub> intermediate photosynthesis can be found in the *Moricandia arvensis*, *M. nitens*, and *M. suffruticosa* monophylum, as well as in *Diplotaxis eruroides*, *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<sub>3</sub> or C<sub>3</sub>-C<sub>4</sub> 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<sub>3</sub> and C<sub>3</sub>-C<sub>4</sub> 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<sub>3</sub>-C<sub>4</sub> genomes (two-way ANOVA,  $P = 0.013$ ).

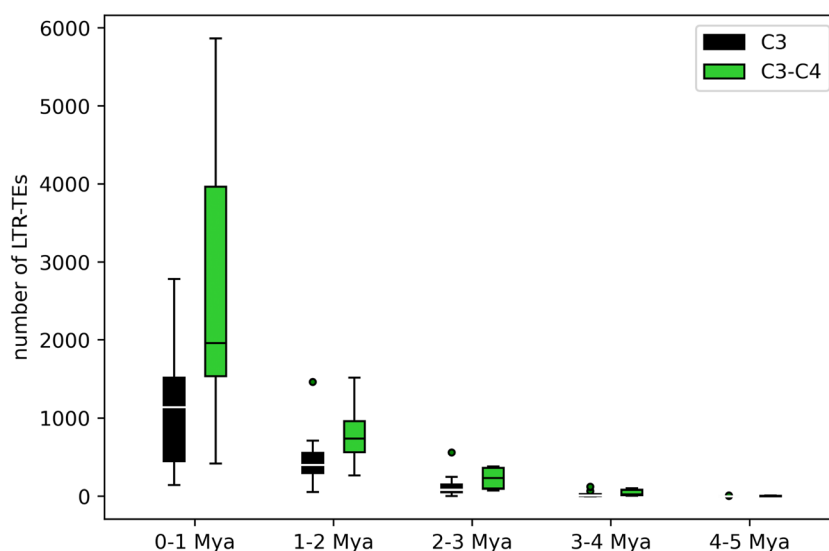
To analyse recent increases of TE activity and their potential roles in the evolution of C<sub>3</sub>-C<sub>4</sub> intermediate photosynthesis, we determined the insertion times of long terminal repeat (LTR)

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<sub>3</sub> and C<sub>3</sub>-C<sub>4</sub> 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<sub>3</sub>-C<sub>4</sub> 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<sub>3</sub> and C<sub>3</sub>-C<sub>4</sub> intermediate genomes

To better understand whether the high abundance of TEs in C<sub>3</sub>-C<sub>4</sub> 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).

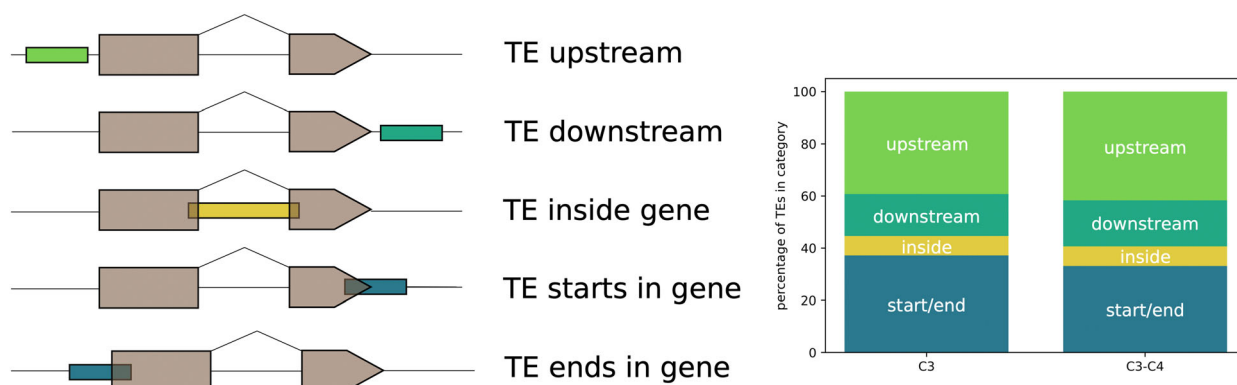
Genes with TEs within the gene model (III) and overlapping TEs (IV and V) might have broken coding sequences and may result from imprecise annotations. Across the selected 11 species, 55,148 TEs were identified to be co-occurring with a protein coding gene in at least one species, whereas 21,643 co-occurring TEs belonged to C<sub>3</sub> and 28,379 co-occurring TEs belonged to C<sub>3</sub>-C<sub>4</sub> species. In both C<sub>3</sub> and C<sub>3</sub>-C<sub>4</sub> intermediate species, over 50% of the TEs co-occurring with genes were



**Fig. 4.** Boxplot of LTR-TE insertion times for eight C<sub>3</sub> and six C<sub>3</sub>-C<sub>4</sub> 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  $1.4 \times 10^{-8}$  substitutions per site per year. Black boxes represent C<sub>3</sub> species, green boxes represent C<sub>3</sub>-C<sub>4</sub> species.



Transposable elements contribute to the establishment of the glycine Triesch, Denton, Bouvier, Buchmann, Reichel-Deland, Guerreiro, Busch, Schlüter, Stich, Kelly & Weber



**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<sub>3</sub>-C<sub>4</sub> intermediate orthologs of these genes exhibited upstream TEs, whereas the C<sub>3</sub> orthologs lacked upstream TEs. Thus, during the evolution of C<sub>3</sub>-C<sub>4</sub>, 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” ( $P = 0.002907$ )

**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	<i>P</i> -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

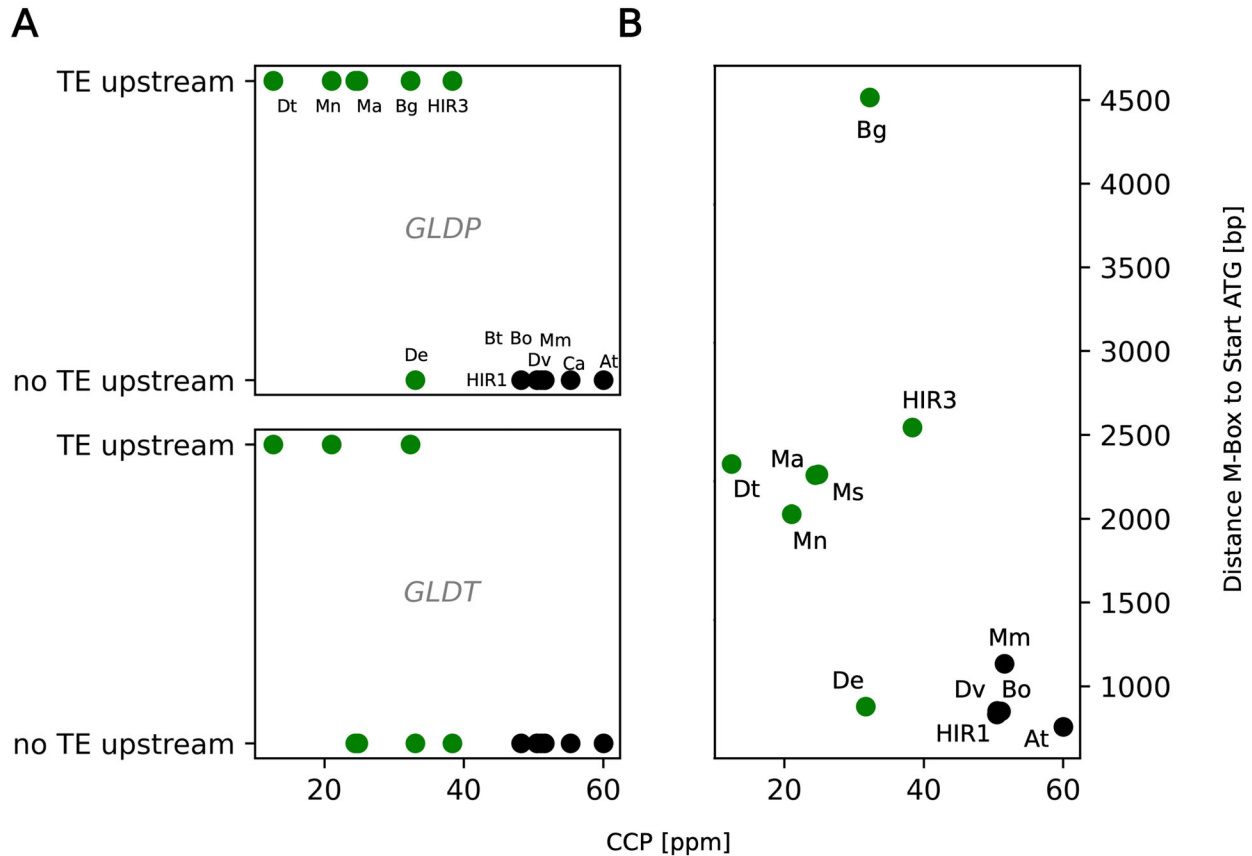
was enriched in the set of genes that co-occur with upstream transposons (Table 2). The occurrence of this *Mercator* bin was increased 38-fold over the background, which is higher than for any other analysed *Mercator* bin (Table 2).

#### The *GLDP1* upstream region shows independent TE insertions in C<sub>3</sub>-C<sub>4</sub> 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 Brassicaceae tribe that contains all known C<sub>3</sub>-C<sub>4</sub> 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<sub>3</sub>-C<sub>4</sub> 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 erucooides*, 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.

<i>Mercator</i> 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. erucooides* (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. erucooides* no insertion could be found.

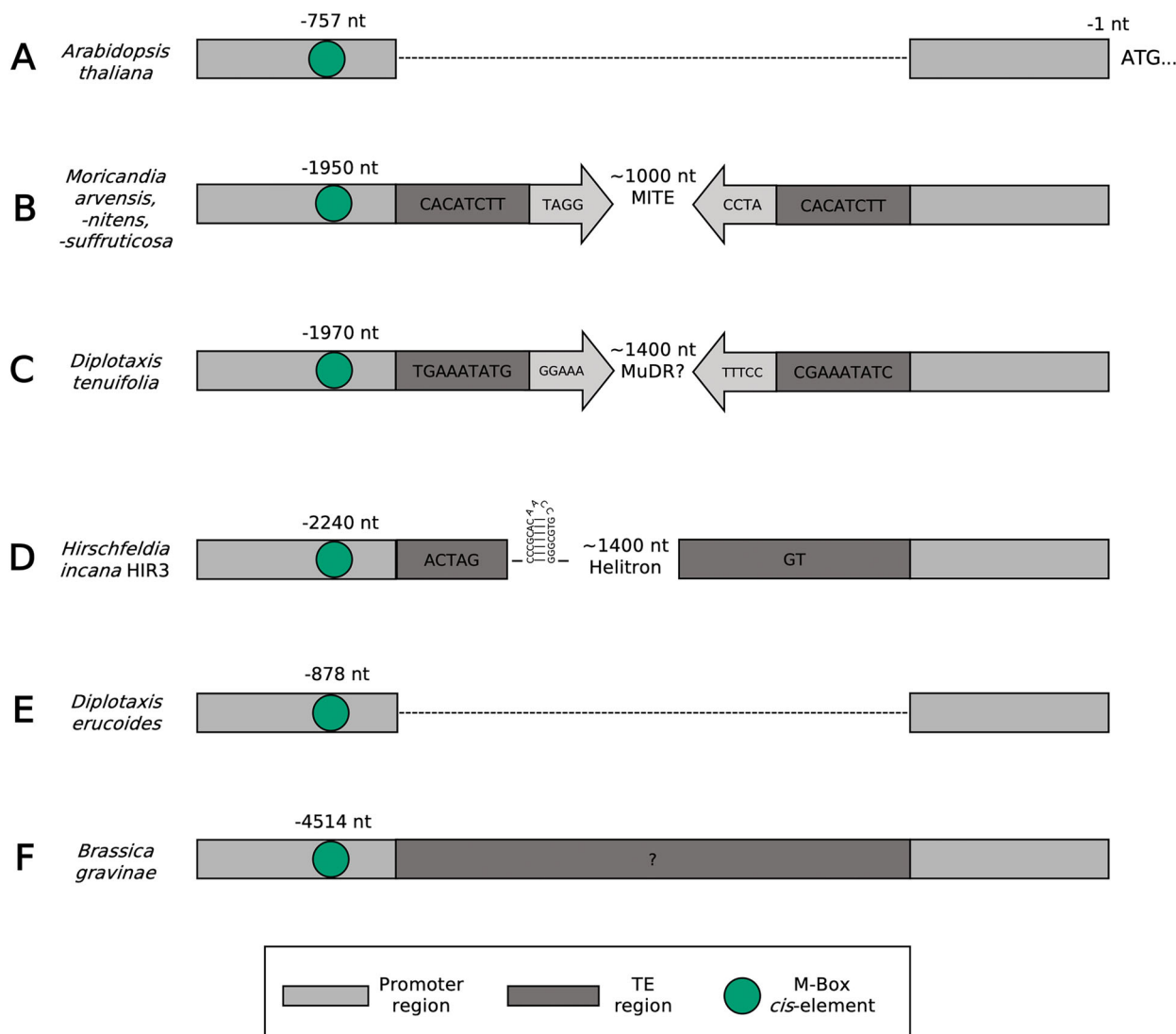
From five analysed  $C_3$ - $C_4$  *GLDP1* promoters, we found a large insertion behind the conserved M-Box in four cases (monophyletic  $C_3$ - $C_4$  intermediate *Moricandia* clade, *Diplotaxis tenuifolia*, *Brassica gravinae* and *Hirschfeldia incana* HIR3; Fig. 6B). Out of these four cases where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

Transposable elements contribute to the establishment of the glycine Triesch, Denton, Bouvier, Buchmann, Reichel-Deland, Guerreiro, Busch, Schlüter, Stich, Kelly & Weber



**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 eruroides* *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 be 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<sub>3</sub>-C<sub>4</sub> intermediate traits can only be found in the Brassicaceae 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<sub>3</sub>-C<sub>4</sub> 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<sub>3</sub>-C<sub>4</sub> intermediate phenotype, we cannot ultimately conclude that LTR transposon bursts contributed to the evolution of the C<sub>3</sub>-C<sub>4</sub> 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<sub>3</sub> and C<sub>3</sub>-C<sub>4</sub> 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<sub>3</sub>-C<sub>4</sub> intermediate traits. Furthermore, based on sequence identity between the C<sub>3</sub>-C<sub>4</sub> 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<sub>3</sub> and C<sub>3</sub>-C<sub>4</sub> intermediate *Moricandia* species of 2–11 million years (Arias *et al.* 2014). Thus, changes in TE content occurred concomitant with the evolution of C<sub>3</sub>-C<sub>4</sub> intermediate photosynthesis and occurred in genes whose expression is required to change for operation of a C<sub>3</sub>-C<sub>4</sub> cycle.

In the descriptive whole-genome view, we observed correlations between TE content and age and the C<sub>3</sub>-C<sub>4</sub> 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<sub>3</sub>-C<sub>4</sub> 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<sub>3</sub>-C<sub>4</sub> intermediate phenotype.

In both C<sub>3</sub> and C<sub>3</sub>-C<sub>4</sub> 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<sub>4</sub> 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<sub>3</sub>-C<sub>4</sub> 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<sub>3</sub>-C<sub>4</sub> 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<sub>3</sub>-C<sub>4</sub> 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<sub>3</sub>-C<sub>4</sub> 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<sub>3</sub>-C<sub>4</sub> evolution across multiple plant lineages (Schulze *et al.* 2013; Schlüter & Weber 2016). In the Brassicaceae tribe, the *GLDP2* copy was lost (Schlüter *et al.* 2017). Additionally, *GLDP1* was reported to be differentially expressed between C<sub>3</sub> and C<sub>3</sub>-C<sub>4</sub> 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<sub>3</sub> species *M. moricandioides* conferred GUS expression to both MC and BSC, whereas the *GLDP1* promoter of the C<sub>3</sub>-C<sub>4</sub> 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<sub>3</sub>-C<sub>4</sub> 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<sub>3</sub>-C<sub>4</sub> 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<sub>3</sub>-C<sub>4</sub> photosynthesis. This hints at a remarkable convergent evolutionary genetic mechanism in C<sub>3</sub>-C<sub>4</sub> evolution. We suggest that the loss of *GLDP2* paved the way for neofunctionalization of the *GLDP1* copy in the Brassicaceae tribe, the only Brassicaceae tribe containing C<sub>3</sub>-C<sub>4</sub> 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<sub>3</sub>-C<sub>4</sub> 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<sub>3</sub> plants (Fig. 6B). However, *D. erucoides* shows C<sub>3</sub>-C<sub>4</sub> intermediate phenotypes (Schlüter *et al.* 2017; Lundgren 2020). We assume that, being an independent evolutionary origin of C<sub>3</sub>-C<sub>4</sub> 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<sub>3</sub>-C<sub>4</sub> 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<sub>3</sub>-C<sub>4</sub> 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<sub>3</sub>-C<sub>4</sub> 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<sub>3</sub>-C<sub>4</sub> intermediate species will pave the way for a better understanding of the C<sub>4</sub> 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/GRK2466, 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](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<sub>2</sub>-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<sub>2</sub> *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 pherel1 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<sub>2</sub> and C<sub>4</sub> photosynthesis in C<sub>3</sub> 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 'nomenclature: 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 C<sub>4</sub> photosynthesis. *BMC Genomics*, **17**, 201.
- Christin P.A., Sage T.L., Edwards E.J., Ogburn R.M., Khoshhaves R., Sage R.F. (2011) Complex evolutionary transitions and the significance of C<sub>3</sub>-C<sub>4</sub> 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 C<sub>3</sub> and C<sub>4</sub> grasses (Poaceae): bundle sheath and mesophyll surface area relationships. *Annals of Botany*, **73**, 241–255.
- Dickinson P.J., Knerová J., Szczepowka 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<sub>2</sub> 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<sub>4</sub> 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<sub>3</sub>-C<sub>4</sub> intermediate photosynthesis in the Brassicaceae 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<sub>4</sub> 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<sub>4</sub> photosynthesis in Cleomeaceae bioRxiv. 2022.07.09.499295.
- Holst F., Bolger A., Günther C., Maß J., Triesch S., Kindel F., Kiel N., Saadat N., Ebenhöf O., Usadel B., Schwacke R., Bolger M., Weber A.P., Denton A.K. (2023) Helixer – de novo prediction of primary eukaryotic gene models combining deep learning and a Hidden Markov 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<sub>3</sub>-C<sub>4</sub> 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<sub>3</sub>, C<sub>4</sub> 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<sub>2</sub> 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<sub>3</sub>–C<sub>4</sub> 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<sub>3</sub>–C<sub>4</sub> 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<sub>3</sub>–C<sub>4</sub> 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.
- Pietzenk 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<sub>3</sub> and C<sub>3</sub>–C<sub>4</sub> 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<sub>4</sub> 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<sub>4</sub> 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<sub>3</sub>–C<sub>4</sub> intermediate *Moricandia* species. *Journal of Experimental Botany*, **68**, 191–206.
- Schlüter U., Weber A.P. (2016) The road to C<sub>4</sub> photosynthesis: evolution of a complex trait via intermediate 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<sub>4</sub> photosynthesis in the genus *Flaveria*: establishment of a photorespiratory CO<sub>2</sub> pump. *The Plant Cell*, **25**, 2522–2535.
- Schulze S., Westhoff P., Gowik U. (2016) Glycine decarboxylase in C<sub>3</sub>, C<sub>4</sub> and C<sub>3</sub>–C<sub>4</sub> 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., Schmick R., Franke 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<sub>4</sub> enzymes in *Gynandropsis gynandra* mesophyll cells. *The Plant Cell*, **28**, 454–465.
- Wiludca C., Schulze S., Gowik U., Engemann S., Koczor M., Streubel M., Bauwe H., Westhoff P. (2012) Regulation of the photorespiratory GLDPA gene in C<sub>4</sub> *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.