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

Sebastian Triesch[1,2], Alisandra K. Denton[1,2], Jan P. Buchmann[2,3], Vanessa Reichel-Deland[1], Ricardo Nuno Ferreira Martins Guerreiro[4], Urte Schlüter[1,2], Benjamin Stich[2,4], 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 Insitute for Biological Data Sciences, Heinrich Heine University Düsseldorf, Germany
4 Insitute for Quantitative Genetics and Genomics of Plants, Heinrich Heine University Düsseldorf, Germany

*** Corresponding author: andreas.weber@uni-duesseldorf.de**

Abstract

C_3-C_4 intermediate photosynthesis has evolved multiple times convergently and independently in the Brassicaceae, although the family lacks *bona fide* C_4 species. Evolution of C_3-C_4 intermediate photosynthesis requires a reconfiguration of gene-regulatory networks, modified spatial expression patterns of multiple genes and ultrastructural adjustments. Mechanisms underpinning the reconfiguration of these networks are mostly unknown. In this study, we use a pan-genomic association approach to identify genomic features that might confer differential gene expression towards C_3-C_4 intermediate traits. We found a strong correlation between transposon insertions in *cis*-regulatory regions and the C_3-C_4 intermediacy trait. Our study revealed 222 orthogroups where 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. The P-protein of the glycine decarboxylase enzyme complex is known to play a crucial role in the establishment of the photorespiratory glycine shuttle. We found independent transposon insertions in the promoter sequences of the gene encoding the P-protein and hypothesize that these insertions lead to a spatial shift in gene expression. Our findings hint at a pivotal role of TEs in the evolution of C_3-C_4 intermediacy, especially in mediating differential gene expression.

Introduction

C_4 photosynthesis convergently and independently 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 and dry areas where leaf internal CO_2 concentrations tend to be low and light conditions are frequently saturating, both promoting the oxygenation ratio of Rubisco (Betti et al. 2016, Sage et al. 2012). Although the trait has evolved convergently across multiple plant lineages, the complexity of anatomical, biochemical, and developmental adaptations makes understanding and modifying C_4 photosynthesis a difficult undertaking.

Plants that exhibit C_3-C_4 intermediate phenotypes are promising research subjects to study the early adaptations 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 that are also observed in C_4 species. They differ from C_4 species in terms of metabolism and do not show a fully integrated C_4 cycle. The C_3-C_4 intermediate trait is characterized by, e.g., a lowered carbon compensation point (CCP), a reduced number of mesophyll cells (MC), chloroplast-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).

One trait that is commonly shared between multiple 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)). It relies on the BSC-specific decarboxylation of photorespiratory glycine, leading to an elevated CO_2 concentration around Rubisco. The increased partial pressure of CO_2 around the site of its fixation leads to a larger proportion of the Rubisco carboxylation reaction, which suppresses photorespiration and decreases the CCP (Kennedy & Laetsch 1974, Monson & Edwards 1984, Schlüter et al. 2017).
It was shown that the BSC-specific decarboxylation of glycine is caused by the differential activity 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 was reported to be only present in BSC mitochondria, but not in MC mitochondria (reviewed in Schulze et al. (2016)). This is a notable example of convergent evolution, since these species belong to the distant families Brassicaceae, Asteraceae and Poaceae, respectively. 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). 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 active (Schulze et al. 2013). In C_3-C_4 intermediate *Moricandia*, the differential expression of *GLDP1* is thought to be mediated by the loss of one gene copy and differential regulation of the other: whereas *GLDP2* is absent in the Brassiceae clade containing the C_3-C_4 intermediate species, *GLDP1* was reported to be differentially expressed by loss of a potential *cis*-element called M-Box. The M-Box confers a basal mesophyll- and bundle sheath expression in the *Arabidopsis thaliana GLDP1* promoter (Adwy et al. 2015, 2019) and was not found in *GLDP1* promoter orthologs of C_3-C_4 intermediate *Moricandia* species (Schlüter et al. 2017). A second *cis*-element, the V-Box, was shown to confer expression to the BSC and is present in all so far analyzed Brassicaceae *GLDP1* promoter sequences (Adwy et al. 2015, 2019). Spatially and temporally differential gene expression is crucial for the evolution of C_3-C_4 intermediate photosynthesis (Hibberd & Covshoff 2010, Reeves et al. 2017). The underlying genomic variation can be mediated by single-nucleotide polymorphisms (SNPs) or structural variation such as insertions or deletions.

Structural variation can be caused by the activity of mobile genetic elements, also called transposable elements (TEs) or transposons. 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. It has been frequently proposed that TEs can contribute greatly to 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 and Davidson put forward that co-option of mobile sequences containing gene regulatory elements can connect genes to the same gene regulatory networks (Britten & Davidson 1971). The co-option of TEs for regulatory purposes is called "exaptation" (Brosius & Gould 1992). With today's vast amount of available genomic data, 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 mode is the exaptation of a *cis*-regulatory element (CRE) from other genes (Fig 1). If the CRE inside a TE is copied from one gene and retained by the other gene, both genes are afterwards controlled by a mutual CRE, thus connected in a gene regulatory network (Fig 1, B). In contrast to this, 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). Another possibility is the *de novo* generation of new CRE by point mutations in degrading 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).
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 connected to endosperm development, e.g., the distribution of the PHEREs1 MADS-box transcription factor binding motifs by *Helitron* transposons in *Arabidopsis thaliana*

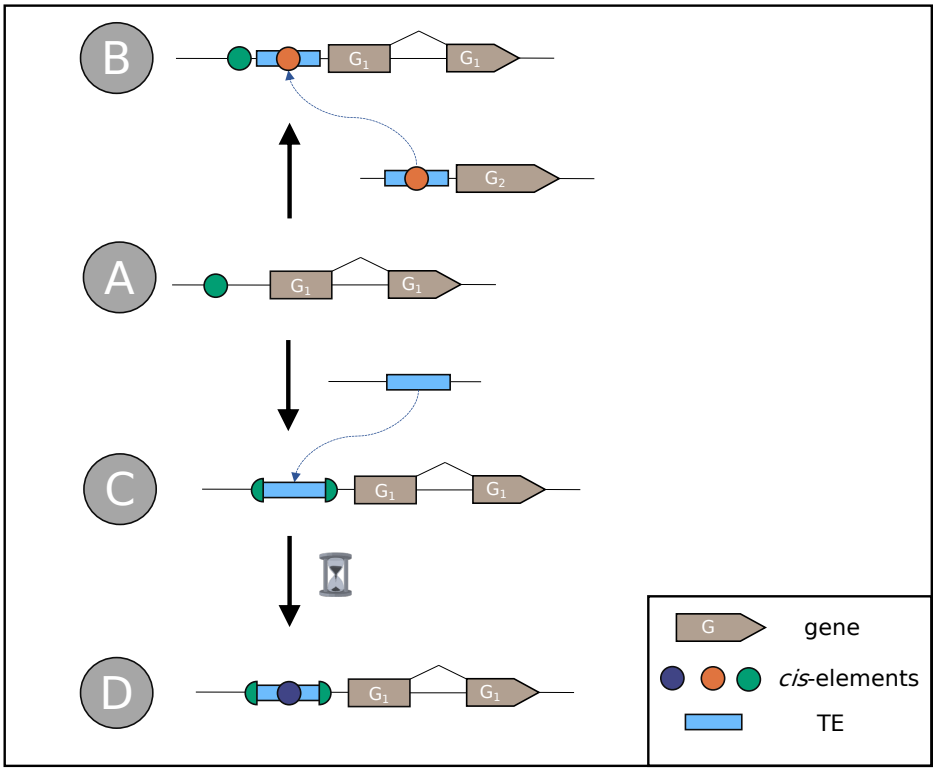

Figure 1. Schematic illustration of gene regulation rewiring by TE exaptation. A: The hypothetical gene G1 is controlled by a CRE (green dot). B: Gene G2 is regulated by a different CRE (orange dot) located within a TE (blue box). Upon transposition of the TE to the upstream region of G1, G1 might co-opt 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.

(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).

Investigating potential roles for TEs in the evolution of C_4 photosynthesis, Cao et al. (2016) analyzed 40 C_4 gene orthologs between rice and maize for the presence of BSC-specific promoter motifs. Of over 1,000 promoter motifs that were differentially distributed between C_3 and C_4 orthologs, more than 60 % were 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. 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.

Here, we focus on TE-mediated exaptation of CREs in C_3-C_4 intermediate Brassicaceae species that mediate decisive steps in the evolution of the photorespiratory glycine shuttle and could play a crucial role in C_4 evolution. We analyzed independent evolutionary events in closely related species from one plant family, the Brassicaceae. The Brassicaceae family 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). Current studies (Guerreiro *et al.*, in preparation) described at least five independent origins of C_3-C_4 intermediate photosynthesis in this family.

To test whether TE insertions might be involved in establishing the C_3-C_4 intermediate photosynthetic trait, we performed a pan-genomic association study to analyze 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 searched for correlations between the presence or absence of upstream co-occurring TEs with the CCP. 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 our analysis, we examined the promoter orthologs of relevant photorespiratory genes in closer detail to reveal the influence of TE insertions towards the establishment of C_3-C_4 photosynthesis traits. Here, we present evidence that the insertion of TEs in *cis-regulatory* regions are associated with the establishment of C_3-C_4 photosynthesis in the Brassicaceae.

Materials and Methods

Genomes and carbon compensation points

The genomes for *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. (in preparation). 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 sequences 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. (in preparation). Data for the phylogenetic tree was obtained from Guerreiro *et al.* (in preparation).

Gene annotation

Consistent structural gene annotations were generated for each species with *Helixer* (Stiehler et al. 2020) using the hybrid convolutional and bidirectional long-short term memory model, `HybridModel`, specifically the trained instance of `land_plant_v0.3_m_0100`. This was followed by post-processing the raw predictions into final primary gene models with Helixer Post (Bolger 2022, personal communication). Default parameters were used except as specified below. During data preparation, the subsequence length was increased to 106920 bp. During inference, overlapping was activated using a core-length of 53460 bp and an offset of 13365 bp. Phase prediction was used. During post processing, a window size of 100 bp, an edge threshold of 0.1, a peak threshold of 0.8, and a minimum coding length of 60 bp were used. The exact code-state can be found at the following commits: *Helixer* bb840b4, *GeenuFF* 1f6cffb and *Helixer Post* 08c6215, while the process is compatible with all releases beginning with v0.3.

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 LTR insertion time was calculated using

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

assuming a neutral mutation rate of $\mu = 1.4 \cdot 10^{-8}$ substitutions per site per year (Cai et al. 2018).

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 scripts are available on GitHub. 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. Genes with TEs within 3000 bp upstream were filtered and collapsed into orthogroups (OGs) using *OrthoFinder* 2.5.4 (Emms & Kelly 2015). For each OG, the presence or absence of an upstream TE was correlated to the CCP using one-way ANOVA. Each OG was functionally annotated using *Mercator4* 4.0 (Schwacke et al. 2019). A randomly chosen representative protein sequence from each OG was uploaded to the *Mercator4* web tool for functional annotation. 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

To screen 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. The five independent origins of C_3-C_4 intermediate photosynthesis can be found in the *Moricandia arvensis*, *-nitens*, and *-suffruticosa* monophylum, as well as in *Diplotaxis erucoides*, *Diplotaxis tenuifolia*, *Brassica gravinae* and *Hirschfeldia incana* HIR3 (Fig. 2).
The species panel exhibits genome sizes ranging from 120 Mbp in *Arabidopsis thaliana* to 677 Mbp in *Moricandia 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 *Arabidopsis thaliana* and 75 % in *Moricandia arvensis*. We observed differences in genome size and TE content also in closely related species, e.g. between *M. arvensis* and *M. moricandioides* or between*B. gravinae* and *D. viminea*. Furthermore, we observed that differences in genome size are mainly due to the different TE content.

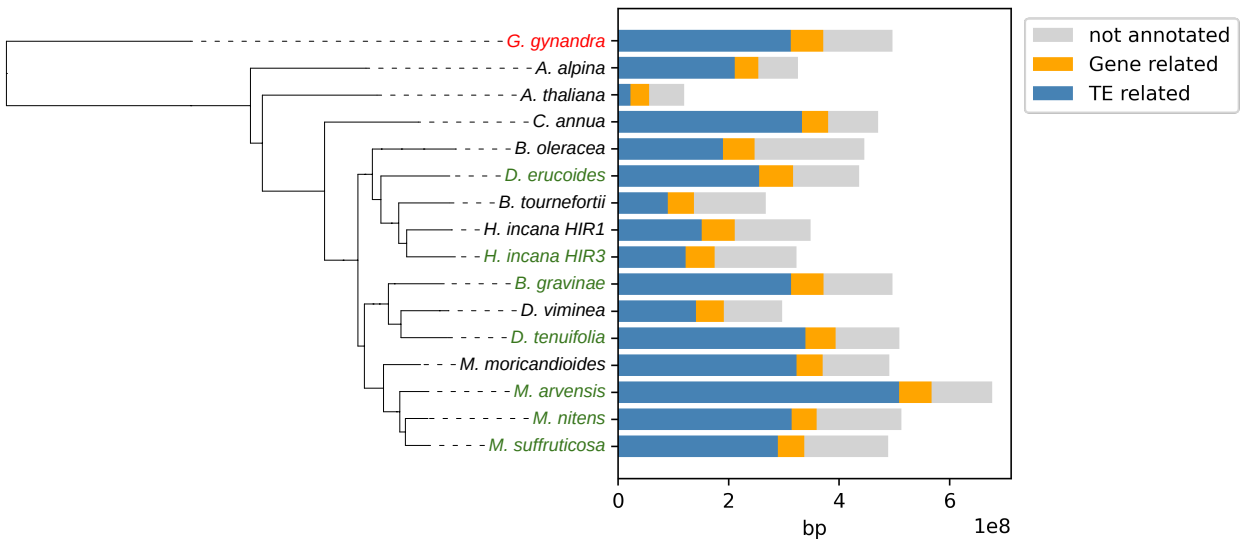

Figure 2. Phylogeny and genomic composition of 15 selected Brassicaceae species and the Cleomaceae outgroup. C_3-C_4 intermediate species are highlighted in green, the C_4 outgroup *G. gynandra* is highlighted in red. TE-related nucleotides are defined as spanning intact and fragmented transposons.

Class I type retrotransposons represented the majority of identified TEs in C_3 and C_3-C_4 species (Fig. 3). Within each analyzed genome, between 60 % and 68 % of all annotated TEs were Class I retrotransposons for C_3 and C_3-C_4 species, respectively. Among all analyzed species, the proportion of TE classes in the

genomes varied greatly (Fig. 3; Supplemental Table S2). In the genomes analyzed here, the TE Class II was dominated by TEs from the *Helitron* group, making up between five and 20 percent 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, $p = 0.013$).

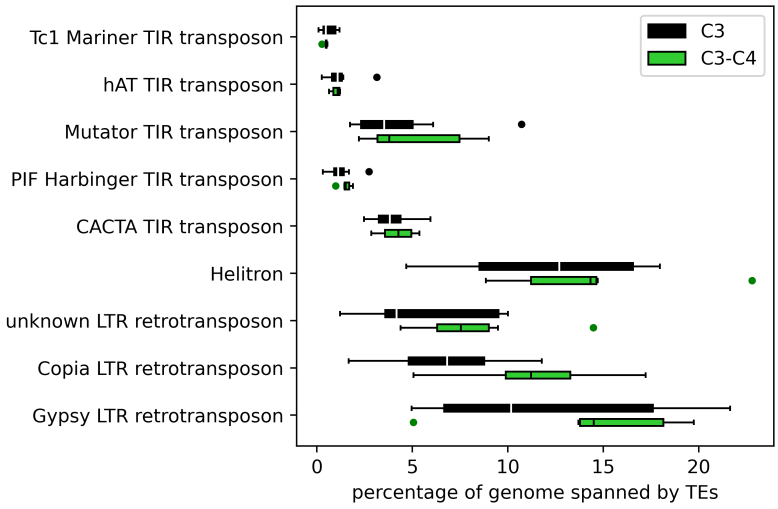

Figure 3. Boxplot indicating the percentage of the genome spanned by nine classes of intact and fragmented TEs in eight C_3 and six C_3-C_4 intermediate species. The y-axis shows the TE classes, the x-axis indicates the fraction of the genome made up by the respective TE class. Black boxes depict C_3 species, green boxes depict C_3-C_4 intermediate species. The boxes start and end at the lower/upper quartiles of the data, the whiskers extend to the upper and lower range of the data. Outliers are indicated by circles. The median is indicated by a white or black bar for C_3 or C_3-C_4 intermediate species, respectively.

To analyze recent increases of TE activity and their potential roles in the evolution of C_3-C_4 intermediate photosynthesis, we determined the LTR insertion times (Fig. 4; Supplemental Table S3). *LTR retriever*, which is the LTR annotation tool of the *EDTA* pipeline, detected LTR transposons down 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 back to a maximum of 4 million years. Both C_3 and C_3-C_4 intermediate species showed mainly the same pattern of LTR bursts. There was an increased frequency for LTR-TEs younger than two million years. The increase was more pronounced for C_3-C_4 intermediate species, which was largely due to 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 (2-way ANOVA, $p = 0.033$).

We next analyzed the differential co-occurrence of TEs with genes. We reduced our species set to five C_3 species and six C_3-C_4 intermediate species based on genome quality and to analyze a comparable number of C_3 and C_3-C_4 species (At, Bo, Bg, De, Dt, Dv, HIR1, HIR3, Ma, Mm, Mn). 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 spans a whole gene (spanning), (IV) the TE starts in a gene (start), or (V) the TE ends in a gene (end).

Genes that were spanned by TEs (III) were mostly open reading frames related to the transposase genes of the respective TEs. Genes with overlapping TEs (IV and V) might have broken coding sequences and may result from imprecise annotations. Across the selected eleven species, 61,092 TEs were identified to be co-occurring with a gene in at least one species, whereas 22,053 co-occurring TEs belonged to C_3 and 39,039 co-occurring TEs belonged to C_3-C_4 species. In both C_3 and C_3-C_4 intermediate species, around 70 % of the TEs co-occurring with genes were located up- or downstream of the gene (Fig. 5). Analyzing

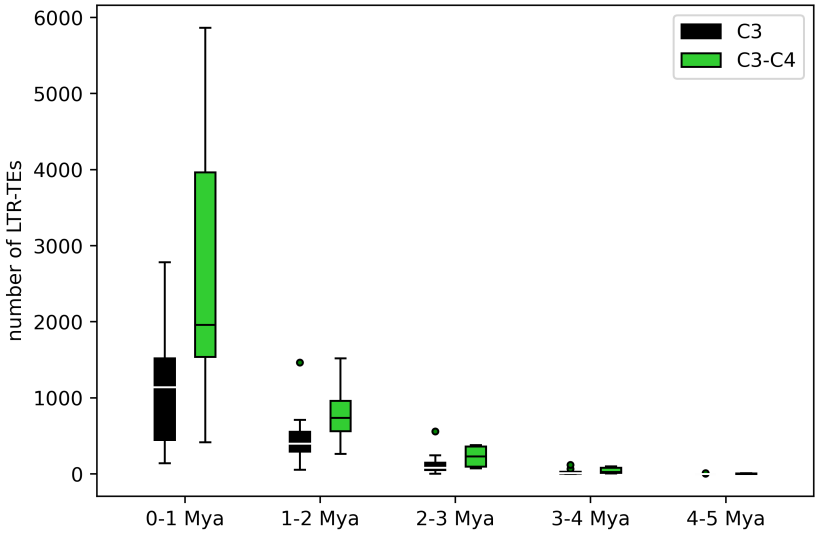

Figure 4. Boxplot of LTR-TE insertion times for eight C_3 and six C_3-C_4 intermediate species. The x-axis shows the insertion time in bins of 1 million years before today (Mya). The y-axis depicts the number of identified LTR-TEs calculated to be inserted within this timeframe. The boxes start and end at the lower/upper quartiles of the data, the whiskers extend to the upper and lower range of the data. Outliers are indicated by circles. The median is indicated by a white or black bar for C_3 or C_3-C_4 intermediate species, respectively. 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.

potentially exaptated CREs, we further analyzed TEs in the up to 3000 bp 5' region of the gene. A total of 22,560 genes with co-occurring upstream TEs were collapsed into 12,087 OGs using *OrthoFinder* (Emms and Kelly, 2015). By this, we filtered out false positive gene models from the *Helixer* annotation and allowed for between-species comparisons. For each of these OGs, one-way ANOVA was employed, correlating the presence or absence of a co-occurring upstream TE with the CCP of the respective species. We identified 222 OGs where the co-occurrence of one of the orthologs with an upstream TE correlated with the CCP ($p \leq 0.05$; Tab. 1, Supplemental Table S4). Among the ten OGs with the lowest p-value were genes involved in photorespiration such as GLYCOLATE OXIDASE or the genes encoding the T- and P-subprotein of the glycine decarboxylase complex (Fig. 6A). Strikingly, in the ten OGs with the lowest p-values, the C_3-C_4 intermediate OGs exhibited upstream TEs, whereas the C_3 OGs lacked upstream OGs. Thus, in the OGs with the lowest p-values, we observed a "gain" in upstream TEs (Fig. 6A).
In the list of OGs with a correlation between the CCP and the presence of upstream TEs, three photorespiratory genes occurred (*GOX*, *GLDP*, *GLDT*). To quantify putative enrichment of certain gene ontologies, each OG was functionally annotated with a *Mercator* bin. Statistical enrichment analysis using Fisher's exact test revealed that the *Mercator* bin "Photosynthesis.Photorespiration" ($p = 0.000904$) was enriched in the set of genes that co-occur with upstream transposons (Tab. 2). The occurrence of this *Mercator* bin was increased twenty-fold over the background, which is higher than for any other analyzed *Mercator* bin (Tab.2; Supplemental Table S5).

From the enriched photorespiratory genes, we selected *GLDP*, the gene encoding the P-protein of the GDC, for a deeper analysis. 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. Furthermore, it is known that the differential expression of

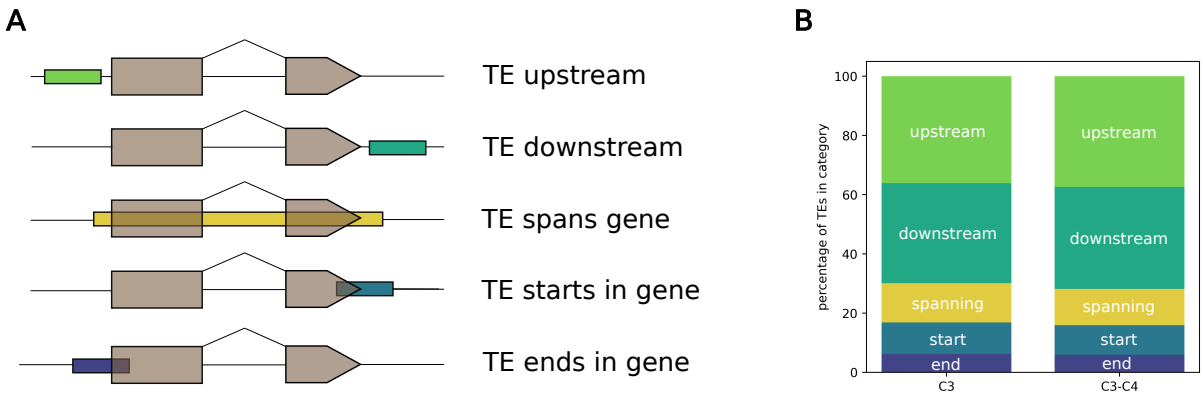

Figure 5. **A:** Different contexts of TEs co-occurring with genes. **B:** Bar charts indicating the fractions of TE co-occurring with genes within five contexts: starting or ending in a gene (start / end), spanning the gene (over) or residing within a 3000 bp window upstream or downstream the gene.

GLDP contributes to the establishment of the photorespiratory glycine shuttle (Monson & Edwards 1984, Rawsthorne et al. 1988, Schulze et al. 2013) and several studies about the underlying regulatory genetics have been conducted (Adwy et al. 2015, 2019, Dickinson et al. 2020, Schulze et al. 2016).
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*, *Moricandia nitens* and *Moricandia 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 the *GLDP1* promoter *MITE* to be 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 mesophyll cell (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 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 analyzed Brassicaceae genomes. 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 *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 analyzed 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, we found evidence for the sequence being a TE in three cases (Fig. 7).

Discussion

Evolution of new traits such as C_3-C_4 and C_4 photosynthesis requires the differential regulation of multiple genes. This includes differential expression between MSC and BSC tissue and 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.

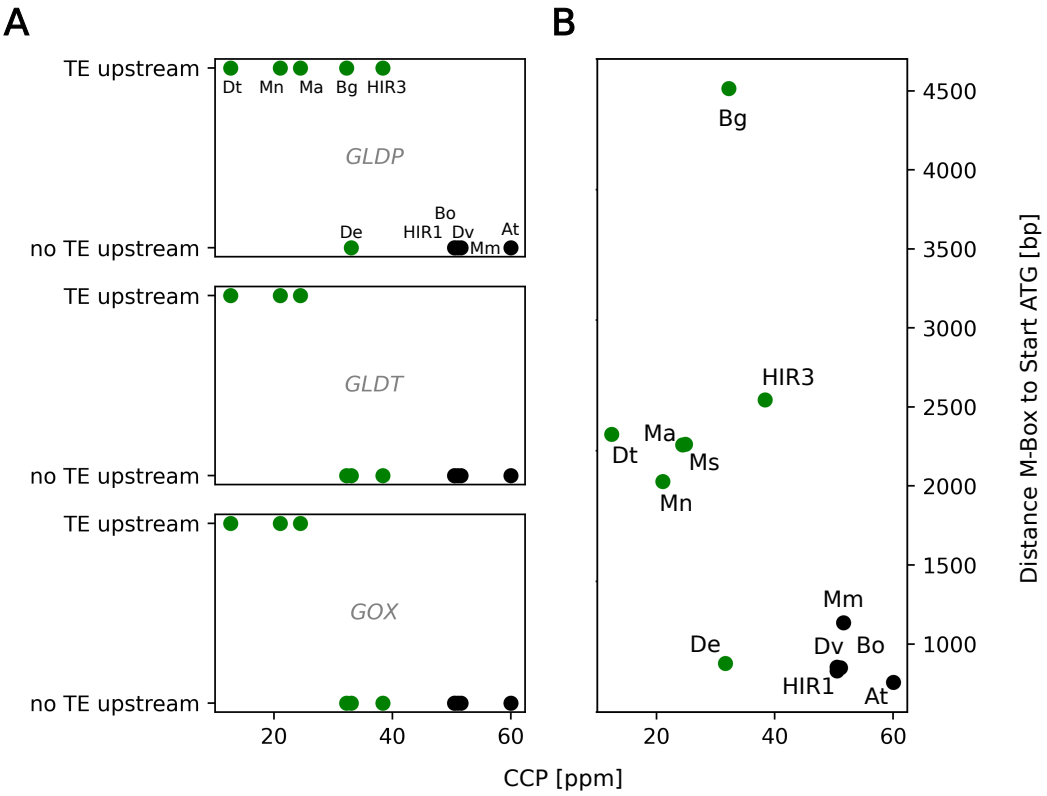

Figure 6. A: Scatter plot for three photorespiratory genes with significant co-associated upstream TEs. The y-axis indicates the presence of an upstream TE (yes/no), the x-axis shows the carbon compensation point. Abbreviations: GLDP/GLDT: P/T-protein of the GLYCINE DECARBOXYLASE COMPLEX; GOX: GLYCOLATE OXIDASE. **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*, 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*

A few *cis*-elements for MC specificity were found, such as the MEM1 motif from the *Flaveria trinervia* phosphoenolpyruvate carboxylase gene (Gowik et al. 2017) or 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 insertion in a target promoter. Furthermore, they 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 has been scarcely investigated. This study presents the *bona fide* first pan-genomic association study correlating the co-occurrence of TEs and genes with the presence of a certain plant phenotype. 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 no correlation of genome size to the presence of the photosynthesis trait was detectable. In our data, different levels of heterozygosity in the sequenced species may confound these findings and genome size estimations have to be handled with

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
transcription factor (WRKY)	AT5G49520	0.0017
homeobox leucine zipper protein	AT1G27050	0.0017
polyketide cyclase/dehydrase/lipid transport superfamily protein	AT5G06440	0.0017
Em-like protein GEA6	AT3G51810	0.0023
regulatory protein (FLZ) of SnRK1 complex	AT5G49120	0.0023
solute transporter (UmamiT)	AT4G15540	0.0023
Protein GLUTAMINE DUMPER 2	AT4G25760	0.0023
aminomethyltransferase component T-protein of glycine cleavage system	AT1G11860	0.0023
glycolate oxidase	AT3G14420	0.0023
glycine dehydrogenase component P-protein of glycine cleavage system	AT4G33010	0.0023

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

Mercator Bin	OGs with p>0.05	OGs with p<0.05	p-value	Odds ratio
Photosynthesis.Photorespiration	8	3	0.000904	20.3
Not assigned.Not annotated	3132	40	0.004306	0.6
Protein biosynthesis.Ribosome biogenesis	122	7	0.009691	3.1
Multi-process regulation.SnRK1-kinase regulation	7	2	0.011104	15.4
Protein homeostasis.Protein quality control	36	3	0.034403	4.5

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 neighboring tribes (Lysak et al. 2009). However, genome size alone does not reflect the complexity of a species, as described by the "C-value paradox" (Thomas 1971).

Next, we analyzed the proportion of TEs in the individual genomes. Our estimation of TE proportions is consistent with previously analyzed Brassicacae genomes (Liu et al. 2020, Mirouze & Vitte 2014) and the *Gynandropsis gynandra* genome (Hoang et al. 2022). We found a significant correlation between the photosynthesis phenotype and the proportion of the genome occupied by TEs in the respective species. Genome size and TE content also vary between closely related species, which may lead to the assumption that these differences are not only due to phylogeny. In all species, 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 of LTR content and age and 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. 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 falls well within the proposed divergence time C_3 and C_3-C_4 intermediate *Moricandia* species of 2-11 million years (Arias, 2014).

The proportion of TEs in the genome can be seen as a proxy for gene duplication events, a process whose role in evolution has been extensively analyzed in the Brassicaceae (Cerbin & Jiang 2018, Oh & Dassanayake 2019, Walden et al. 2020). Evidence for gene duplication events towards C_4 photosynthesis exist, but it is yet unclear which events in the gradual evolutionary pathway along the C_3-C_4 intermediate stages may require

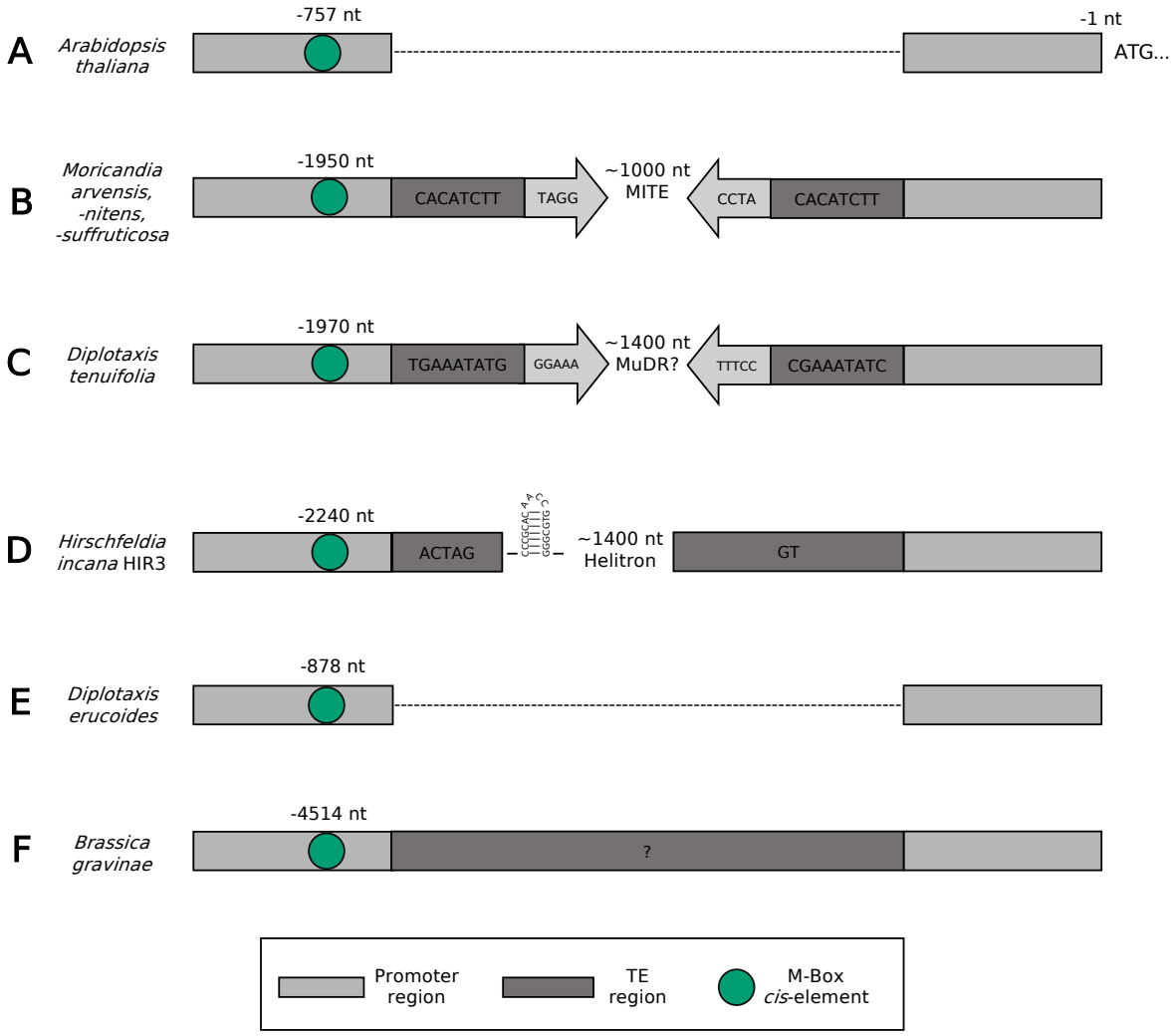

Figure 7. Schematic representation of the *GLDP1* promoter region. "ATG..." depicts the start site of the *GLDP1* gene. Dark grey boxes represent characteristic TE sites such as target site duplications or the *Helitron* insertion sites. Grey arrows depict terminal inverted repeat motifs. The M-Box motif 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 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**).

sub- or neofunctionalized genes (Emms et al. 2016, Gowik et al. 2011, Huang et al. 2021).

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 analyze 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 70 % of the analyzed co-occurring TEs were upstream or downstream of the respective co-occurring gene or spanning the gene. This is biologically plausible, since TEs crossing gene borders may disturb gene function and intergenic regions can harbor transposable elements (Buchmann et al. 2012). Therefore we are confident that our association study, relying on *in silico* predicted genes and TEs, is highly robust.

Differential gene regulation mediated by variation in upstream regions was shown to be a driver of C_4 traits in multiple, well documented cases (Adwy et al. 2015, Gowik et al. 2017, Williams et al. 2015, Wiludda et al. 2012). Our analysis revealed 222 OGs 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 (3 out of 11 OGs with $p<0.05$ for "Photosynthesis. Photorespiration"; 2 out of 9 OGs with $p<0.05$ for "Multi-process regulation.SnRK1 regulation"; see Table 2). We hypothesize that TE retention upstream of these OGs has functional consequences such as differential gene expression, putatively due to the co-option of new- or suppression of existing CREs. Strikingly, the set of genes with significant upstream TEs contains multiple genes involved in photorespiration, such as those encoding GLYCOLATE OXIDASE (GOX) as well as 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 one hint that TEs play a significant role in the introduction of the glycine shuttle.

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 the V-Box, a second promoter element shortly before the At*GLDP1* start site, confers BSC expression (Adwy et al. 2015). 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. Contrary to this, we found the M-Box sequence in all our analyzed *GLDP1* promoter variants. However, we found the M-Box to be shifted 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. This shift may have led to the M-Box being overlooked in previous studies.

Based on the findings by Adwy et al. (2019) we hypothesize that not the loss of the M-Box conferred differential *GLDP1* expression but rather the upstream shift of the element by insertion of a TE. 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 hypothesize that the loss of *GLDP2* paved the way for subfunctionalization of the *GLDP1* copy in the Brassiceae tribe, the only Brassicaceae tribe containing C_3-C_4 intermediate species. This subfunctionalization was mediated by the insertion of a TE in the promoter, suppressing the M-Box element and shifting the *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 hypothesize 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 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.

Based on a whole-genome view of TE density and gene-TE associations, our study revealed a particular scenario in which independent TE insertions contributed to the convergent evolution of a plant trait. Our study highlights the importance of TEs in the evolution of plant traits like C_3-C_4 intermediate photosynthesis. Differential *GLDP* expression is one of the most prominent drivers establishing the glycine shuttle. Based on *in silico* data, the proposed genetic mechanism of differential *GLDP1* regulation by a TE-mediated upstream shift of the M-Box must be verified in experimental work. Looking at the 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 is 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.

Author contributions

A.P.M.W., B.S. and U.S. designed and coordinated the project, S.T. designed and performed all analyses, A.K.D., R.N.F.M.G. and B.S. advised on statistical testing, A.K.D. performed gene annotations using *Helixer*, all authors contributed to writing and accepted the manuscript.

References

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).

Cerbin, S. & Jiang, N. (2018), 'Duplication of host genes by transposable elements', *Current Opinion in Genetics and Development* **49**, 63–69.

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., Kneřová, J., Szecówka, M., Stevenson, S. R., Burgess, S. J., Mulvey, H., Bågman, 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.

Emms, D. M., Covshoff, S., Hibberd, J. M. & Kelly, S. (2016), 'Independent and Parallel Evolution of New Genes by Gene Duplication in Two Origins of C4 Photosynthesis Provides New Insight into the Mechanism of Phloem Loading in C4 Species', *Molecular Biology and Evolution* **33**(7), 1796–1806.

Emms, D. M. & Kelly, S. (2015), 'OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy', *Genome Biology* **16**(1).

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

Gowik, U., Bräutigam, A., Weber, K. L., Weber, A. P. & Westhoff, P. (2011), 'Evolution of C4 photosynthesis in the genus flaveria: How many and which genes does it take to make C4?', *Plant Cell* **23**(6), 2087–2105.

Gowik, U., Schulze, S., Saladié, 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.

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.

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.

Huang, C. F., Liu, W. Y., Lu, M. Y. J., Chen, Y. H., Ku, M. S. & Li, W. H. (2021), 'Whole-Genome Duplication Facilitated the Evolution of C4Photosynthesis in Gynandropsis gynandra', *Molecular Biology and Evolution* **38**(11), 4715–4731.

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).

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.

Oh, D. H. & Dassanayake, M. (2019), 'Landscape of gene transposition-duplication within the Brassicaceae family', *DNA Research* **26**(1), 21–36.

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., Grangé-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.

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.

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.

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

Thomas, C. A. (1971), 'The genetic organization of chromosomes.', *Annual review of genetics* **5**, 237–256.

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.