Insights and Evolutionary mechanisms of C_3-C_4 intermediate Photosynthesis in Brassicaceae: A Genomic and Physiological approach

Inaugural dissertation

for the attainment of the title of doctor
in the Faculty of Mathematics and Natural Sciences
at the Heinrich Heine University Düsseldorf

presented by

Ricardo Nuno Ferreira Martins Guerreiro
from Porto, Portugal

Düsseldorf, March 2025

from the Institute for Quantitative Genetics and Genomics
at Heinrich Heine University Düsseldorf

Published by permission of the
Faculty of Mathematics and Natural Sciences at
Heinrich Heine University Düsseldorf

Supervisor: Prof. Dr. Benjamin Stich
Co-supervisor: Prof. Dr. Andreas Weber

Date of the oral examination: 11 September, 2025

Summary (English)

This thesis presents a comprehensive genomic and physiological study of C_3-C_4 intermediate photosynthesis within the Brassicaceae family. A new panel of species, including 19 novel genome assemblies and both C_3-C_4 types and close C_3 relatives, was established and meticulously characterized for traits such as carbon compensation point, water-use efficiency, vein density, and other characteristics reflective of the C_3-C_4 phenotype. C_3-C_4 species in Brassicaceae exhibited a wide range of carbon compensation points but lacked C_4-like traits such as increased venation or high concentrations of C_4 shuttle metabolites. Metabolic profiles were diverse, with glycine and serine universally present in all C_3-C_4 species, while other metabolites, important for nitrogen balance, varied between species. A unifying feature across all C_3-C_4 Brassicaceae species was centripetal organelle accumulation in bundle sheath cells, an anatomical adaptation that likely enhances CO_2 uptake by minimizing diffusion losses. Notably, C_3-C_4 Brassicaceae do not exhibit significant water-use or nitrogen efficiencies under current atmospheric conditions but do under reduced carbon atmosphere typical of the past.

The genomic assemblies produced include high-quality gene annotations and identified orthology across species, enabling robust comparative analyses. Joint physiological measurements and phylogenetic analysis revealed a new independent origin of the C_3-C_4 trait in a *Hirschfeldia incana* genotype, marking another example of convergent evolution in this family and highlighting its relevance to commercially important species. A critical discovery involved the differential expression of the *GLDP* gene, essential for initiating C_3-C_4 evolution. Genomic comparisons identified a key promoter element whose increased distance from the *GLDP* gene correlates with the C_3-C_4 phenotype, challenging prior assumptions. Alterations in the upstream region were associated with transcription factors, emphasizing their role in plant evolution. These methods can be extended to other genes in the future, leveraging the published panel of species to uncover further insights.

The findings have significant implications for understanding photosynthetic evolution and offer a valuable foundation for improving crop efficiency and resilience. The panel of Brassicaceae species serves as a critical resource for ongoing research. C_3-C_4 traits provide not only a steppingstone toward full C_4 development but also a potential standalone improvement for crops in unstable environments, offering a promising avenue for agricultural sustainability.

Summary (German)

Diese Dissertation präsentiert eine umfassende genomische und physiologische Untersuchung der C_3-C_4-Intermediärphotosynthese innerhalb der Familie der Brassicaceae. Eine neue Gruppe von Arten, einschließlich 19 neu erstellter Genomassemblierungen sowie C_3-C_4-Typen und eng verwandter C_3-Arten, wurde etabliert und detailliert für Merkmale wie den Kohlenstoffkompensationspunkt, Wassernutzungseffizienz, Blattaderdichte und andere für den C_3-C_4-Phänotyp typische Eigenschaften charakterisiert. C_3-C_4-Arten der Brassicaceae zeigten eine breite Spanne von Kohlenstoffkompensationspunkten, wiesen jedoch keine C_4-ähnlichen Merkmale wie erhöhte Blattaderdichte oder hohe Konzentrationen von C_4-Shuttle-Metaboliten auf. Die metabolischen Profile waren vielfältig, wobei Glycin und Serin in allen C_3-C_4-Arten universell vorhanden waren, während andere Metaboliten, die für den Stickstoffhaushalt wichtig sind, zwischen den Arten variierten. Ein gemeinsames Merkmal aller C_3-C_4-Brassicaceae-Arten war die zentripetale Ansammlung von Organellen in den Bündelscheidenzellen, eine anatomische Anpassung, die wahrscheinlich die CO_2-Aufnahme durch Minimierung von Diffusionsverlusten verbessert. Bemerkenswerterweise weisen C_3-C_4-Brassicaceae unter den aktuellen atmosphärischen Bedingungen keine signifikante Wassernutzungs- oder Stickstoffeffizienz auf.

Die erstellten Genomassemblierungen umfassen qualitativ hochwertige Genannotationen und identifizierten Orthologie zwischen den Arten, was robuste vergleichende Analysen ermöglicht. Gemeinsame physiologische Messungen und phylogenetische Analysen enthüllten einen neuen unabhängigen Ursprung des C_3-C_4-Merkmals in einem Genotyp von *Hirschfeldia incana*, ein weiteres Beispiel für konvergente Evolution in dieser Familie, das die Relevanz für kommerziell wichtige Arten hervorhebt. Eine zentrale Entdeckung war die differentielle Expression des *GLDP*-Gens, die für die Initiierung der C_3-C_4-Evolution wesentlich ist. Genomvergleiche identifizierten ein Schlüsselelement im Promotorbereich, dessen vergrößerter Abstand zum *GLDP*-Gen mit dem C_3-C_4-Phänotyp korreliert und frühere Annahmen in Frage stellt. Veränderungen im upstream-Bereich wurden mit Transkriptionsfaktoren in Verbindung gebracht, was ihre Rolle in der Pflanzenevolution unterstreicht. Diese Methoden können in Zukunft auf andere Gene angewendet werden, um weitere Erkenntnisse aus dem veröffentlichten Artenpanel zu gewinnen.

Die Ergebnisse haben bedeutende Implikationen für das Verständnis der photosynthetischen Evolution und bieten eine wertvolle Grundlage zur Verbesserung der Effizienz und Resilienz

von Nutzpflanzen. Das Panel von Brassicaceae-Arten dient als wichtige Ressource für zukünftige Forschungen. C_3-C_4-Merkmale bieten nicht nur einen Zwischenschritt zur vollständigen Entwicklung der C_4-Photosynthese, sondern auch eine potenzielle eigenständige Verbesserung für Pflanzen in instabilen Umgebungen und eröffnen vielversprechende Perspektiven für eine nachhaltige Landwirtschaft.

Glossary

2PG – 2-P-glycolate

3-PGA – 3-phosphoglycerate

ATP – Adenosine triphosphate

BS – Bundle Sheath

CCP – Carbon Compensation Point

CO_2 – Carbon Dioxide

NADPH – Nicotinamide adenine dinucleotide phosphate

NH_3 – Ammonia

GDC – Glycine decarboxylase Complex

GLDP – P-subunit of Glycine decarboxylase Complex

GLDH – H-subunit of Glycine decarboxylase Complex

GLDL – L-subunit of Glycine decarboxylase Complex

GLDT – T-subunit of Glycine decarboxylase Complex

GWAS – Genome Wide Association Mapping

MDH-NADP – Malate dehydrogenase

ME-NADP – Malic enzyme

OAA – Oxaloacetate

PAM – Phylogenetic Association Mapping

PEP – Phosphoenolpyruvate

PEPC – PEP-carboxylase

PPDK – Orthophosphate dikinase

QTL – Quantitative Trait Loci

RuBP – Ribulose-1,5-bisphosphate

RuBisCO – Ribulose-1,5-bisphosphate carboxylase

SNP – Single Nucleotide Polymorphisms

SV – Structural Variation

TADs – Topologically Associated Domains

TE – Transposable Elements

TF – Transcription Factors

AspAT – Aspartate Aminotransferase

DiT2 – Dicarboxylate transporter 2

PPT – Phosphoenolpyruvate/phosphate translocator

Table of contents

Summary (English) I
Summary (German) II
Glossary IV
Table of contents V
List of publications VII
Introduction 1
Photosynthesis and RuBisCO 1
Photorespiration 2
Evolution of Kranz anatomy and carbon concentration mechanisms 2
C_3-C_4 intermediate photosynthesis 3
C_2 glycine shuttle 3
Metabolic background 5
C_3-C_4 as intermediary stage or as evolutionary endpoint? 5
Relevance for agriculture 6
Crop breeding and engineering 6
Genomes and mechanisms of genetic regulation 8
Gene regulatory elements and transcription factors 8
Importance of non-coding genomic variation for gene expression 8
Transposons as sources of genomic variation 9
Foundations for genomics research 10
Phylogenetics of C_3-C_4 photosynthesis 11
C_3-C_4 in Brassicaceae 12
Aims of Thesis 14
Publications 15
1. A genomic panel for studying C_3-C_4 intermediate photosynthesis in the Brassiceae tribe 15
2. Brassicaceae display variation in efficiency of photorespiratory carbon-recapturing mechanisms 33
3. Transposable elements contribute to the establishment of the glycine shuttle in Brassicaceae species 53
Discussion 66
Bioinformatical approaches to study C_3-C_4 in Brassicaceae 66
Genome assembly and annotation 66
Orthology and phylogenetic relations 68
Anatomy and physiology of C_3-C_4 photosynthesis in Brassicaceae 69
Common characteristics of C_3-C_4 70
Advantages and evolution of C_3-C_4 71
Diverse metabolic backgrounds 72
Genetic regulation mechanisms 74
Future avenues for C_3-C_4 research in *Brassicaceae* 76
Directions of research 76

Phylogenetic association mapping 78
Downstream genomic research 79
Crop engineering 80
Conclusion 82
Bibliography 83
Acknowledgments 109

List of publications

Publication 1: Guerreiro, R., Bonthala, V. S., Schlüter, U., Hoang, N. V., Triesch, S., Schranz, M. E., Weber, A.P.M. & Stich, B. (2023). A genomic panel for studying C_3-C_4 intermediate photosynthesis in the Brassiceae tribe. *Plant, Cell & Environment*, *46*(11), 3611-3627.

Publication 2: Schlüter, U., Bouvier, J. W., Guerreiro, R., Malisic, M., Kontny, C., Westhoff, P., Stich, B., & Weber, A. P. M. (2023). Brassicaceae display variation in efficiency of photorespiratory carbon recapturing mechanisms. *Journal of Experimental Botany*, erad250-erad250.

Publication 3: Triesch, S., Denton, A.K., Bouvier, J.W., Buchmann, J.P., Reichel-Deland, V., Guerreiro, R., Busch, N., Schlüter, U., Stich, B., Kelly, S., Weber, A.P.M. (2023). Transposable elements contribute to the establishment of the glycine shuttle in Brassicaceae species. *Plant Biology, 26*, 270–281.

Supplementary Publication 1: Freire*, R.., Weisweiler*, M., Guerreiro*, R., Baig, N., Hüttel, B., Obeng-Hinneh, E., E., Renner, J., Hartje, S., Muders, K., Truberg, B., Rosen, A., Prigge, V., Bruckmüller, J., Lübeck, J. & Stich, B. (2021). Chromosome-scale reference genome assembly of a diploid potato clone derived from an elite variety. *G3*, *11*(12), jkab330.
*Equal first author contribution

Supplementary Publication 2: Schmidt, M., Guerreiro, R., Baig, N., Habekuß, A., Will, T., Ruckwied, B., & Stich, B. (2024). Fine mapping a QTL for BYDV-PAV resistance in maize. *Theoretical and Applied Genetics*, *137*(7), 163.

Supplementary Publication 3: Gerisch, N., Guerreiro, R., Wespel, F., & Stich, B. (2022). Development of a near-infrared spectroscopy calibration for Hagberg falling number assessment of barley (*Hordeum vulgare*): A comparison of methods. *Plant Breeding*, *141*(3), 355-365.

Introduction

Photosynthesis and RuBisCO

Photosynthesis is a fundamental process in the plant realm and for life on the planet (Arnon, 1959). It is the process by which plants and algae convert water, sunlight energy and atmospheric CO_2 into chemical energy in the form of sugar, the basis of the food chain for most living organisms (Arnon, 1951). Plant cells do this with their specialized organelles, the chloroplasts, where photon energy is used to break water molecules (photolysis) into a free proton (H^+ ion), a free electron and oxygen (O_2) (Barber, 2002). As a result, a gradient of protons (chemiosmosis) and a flow of excited electrons are established, which culminate in the creation of ATP (adenosine triphosphate) and NADPH (nicotinamide adenine dinucleotide phosphate) (Mitchell, 1961). These two highly reactive energy-rich molecules are utilized in a downstream pathway, named the Calvin cycle, to build more stable and energy-dense carbohydrates like sucrose (Calvin & Benson, 1948). The Calvin cycle requires the input of atmospheric carbon, captured by the RuBisCO enzyme (Weissbach et al., 1956). Standing for ribulose-1,5-bisphosphate carboxylase, RuBisCO is an essential enzyme for carbon fixation and is so ubiquitous that it is often considered the most abundant protein on the planet (Bar-On & Milo, 2019). Given the diversity of physiology and efficiency regarding the context of RuBisCO function, the identification of underlying genetic diversity may one day inform the transformation of commercial crop species to enhance productivity, stress resilience, and water conservation (Miglani et al., 2021). RuBisCO commonly uses CO_2 to catalyse the carboxylation of ribulose-1,5-bisphosphate (RuBP), a five-carbon molecule, producing two molecules of 3-phosphoglycerate (3-PGA), three-carbon compounds, and releasing Oxygen as a byproduct (Parry et al., 2003). The three-carbon outputs of carbon fixation in this kind of common photosynthesis type earn it the name of C_3 photosynthesis. Despite its predominance, RuBisCO has inherent limitations: besides a relatively low catalytic rate, it can counter-productively bind Oxygen instead of CO_2, oxygenating RuBP and breaking it into a molecule of 3-PGA and one molecule of 2-P-glycolate (2PG) (Bowes et al., 1971). The latter molecule is a competitive inhibitor of other enzymes in the Calvin cycle (Flügel et al., 2017) and it needs to be metabolized to avoid toxicity, which happens through a process called photorespiration (Bauwe et al., 2010).

Photorespiration

In order to recycle the toxic 2PG compound, photorespiration requires the expense of ATP energy and several metabolization steps, acting contrary to the Calvin cycle (Bauwe et al., 2010). Photorespiration originated as a partner of oxygenic photosynthesis billions of years ago, being composed of enzymes of cyanobacterial and proteobacterial origin (Bauwe et al., 2010). It initially converts the toxic 2PG into glycolate in chloroplasts and transports it into a peroxisome, where it is further metabolized into glyoxylate and aminated into glycine (Bauwe et al., 2010). In the mitochondria, pairs of glycine molecules are converted by coordinated reactions of the glycine decarboxylase complex (GDC) and the serine hydroxymethyl transferase into one molecule of serine, releasing CO_2 and NH_3 (for recent reviews see: Eisenhut et al., 2019; Timm & Hagemann, 2020). Through further reactions taking place in the peroxisome and plastid, serine is deaminated into hydroxypyruate, then metabolized into glycerate and finally converted into 3-PGA, the metabolite that can enter the Calvin cycle (Eisenhut et al., 2019). All these steps may reduce the efficiency of photosynthesis by up to 25% depending on abiotic conditions (Bauwe et al., 2010). As photorespiration is highly connected to cellular central metabolism (Bauwe et al., 2012), it is not a purely negative cycle, synthesizing new amino acids and one-carbon units (Busch et al., 2018), as well as both generating and consuming H_2O_2 that is involved in redox signaling and redox homeostasis (Igamberdiev et al., 2001; Foyer et al., 2009). However, it can have a detrimental impact on plant growth especially under hot and dry conditions (Bauwe et al., 2010). In fact, RuBisCO's affinity for O_2 increases at higher temperatures and lower CO_2 concentrations (Salvucci & Crafts-Brandner, 2004), with the addictive factor that leaves' stomata may close under water scarcity, further consuming the CO_2 levels inside the leaf through photosynthesis (Cornic, 2000). These effects will henceforward be referred to as the photorespiratory problem.

Evolution of Kranz anatomy and carbon concentration mechanisms

As a solution to avoid the oxygenation of RuBP, some plant lineages have evolved leaf anatomy and physiology that isolate the Calvin cycle and RuBisCO in specialized cells, where carbon is artificially concentrated through an intermediary transportation mechanism (Hatch, 1987). These specialized cells occur around the vein bundles of the leaf and are hence called Bundle Sheath (BS) cells (Griffiths et al., 2013). This refined leaf anatomy, also called Kranz anatomy, is an astonishing example of convergent evolution, having evolved independently in at least 60

plant lineages (Sage et al., 2011a; Edwards & Voznesenskaya, 2011), though it varies in details (Muhaidat et al., 2007). While the prominence of BS cells is common across evolutionary lineages that avoid photorespiration, the transportation pathway used to increase carbon concentration in BS cells varies in composition and complexity. The complex C_4 pathway is the most widely known, where initially carbon is fixed on phosphoenolpyruvate (PEP), via the PEP-carboxylase (PEPC). The output of this reaction is a four-carbon compound called oxaloacetate (OAA), whence the name C_4 photosynthesis comes from. OAA is further converted into malate or aspartate and then transported to the BS cells, where it releases carbon through decarboxylation via NADP-malic enzyme or NADP-aspartate aminotransferase (Sage, 2004). C_4 evolution from C_3 photosynthesis requires intricate modifications such as structural adaptations related to cell size and morphology, adaptation of enzymes for novel biochemical processes, and spatial reorganization of proteins and organelles (Sage, 2004; Heckmann et al., 2013). In a genetic engineering view, it certainly requires more than a few genetic mutations to achieve. However, there are extant plants that exhibit intermediate traits, including a simpler proto-Kranz anatomy (Sage et al., 2012; Sage et al., 2013), or simpler carbon shuttle mechanisms (Rawsthorne et al., 1992; Mallmann et al., 2014). Measurable parameters such as the CO_2 fixation rate, CO_2 compensation point (CCP) or the organelle accumulation in bundle sheath cells usually show intermediate values between C_3 and C_4 plants (Krenzer, 1975; Ku et al., 1991; McKown & Dengler, 2007; Muhaidat et al.; 2011). These plant lineages are classified under the umbrella term "C_3-C_4 intermediate photosynthesis", but they represent distinct variations along the cline between C_3 and C_4 photosynthesis (Monson et al., 1984; Heckmann et al., 2013).

C_3-C_4 intermediate photosynthesis

C_2 glycine shuttle

Instead of the C_4 shuttle, C_3-C_4 plants shuttle glycine from the mesophyll to BS cells (Edwards and Ku, 1987; Rawsthorne et al., 1992; Sage et al., 2014), where they maintain active glycine decarboxylation for increased CO_2 levels. This shuttle mechanism is also called C_2 shuttle, owing to the 2-carbon composition of glycine, or photorespiratory CO_2 pump, as it involves rearranged enzymes from photorespiration (Sage et al., 2014; Schulze et al., 2016). Glycine itself is produced in photorespiration during the degradation of 2PG (Eisenhut et al., 2019), and

its decarboxylation is mediated by the glycine decarboxylase complex (GDC), a multi-enzyme comprised of 4 proteins (GLDP, GLDH, GLDL, GLDT) (Kikuchi & Hiraga, 1982). GDC is located in the mitochondria, and it cleaves photorespiratory glycine into serine, CO_2, ammonia (NH_3) and NADH (Schulze et al., 2016). The establishment of the C_2 shuttle is comparatively more straightforward than the full C_4 cycle, as it can emerge simply from differential expression of GDC in the BS cells (Monson et al., 1984; Rawsthorne et al., 1988a; Sage et al., 2014). In fact, the differential expression of the P-subunit of GDC is enough to limit the functionality of GDC to BS cells (Rawsthorne et al., 1988a, b; Schulze et al., 2016). The exclusive glycine decarboxylation in BS cells then promotes entry of more photorespiratory glycine from the Mesophyll through diffusion, acting as a glycine sink (Rawsthorne, 1992; Bräutigam & Gowik, 2016). Evolutionary pressure should then occur to favour larger cell sizes and higher concentration of chloroplasts and mitochondria in BS cells, translating to direct efficiency gains (Schlüter & Weber, 2016; Lundgren, 2021). Other important photorespiratory and photosynthetic enzymes, including the other subunits of GDC, RuBisCO, serine hydroxymethyltransferase, glycolate oxidase, have been found both in BS and Mesophyll cells of C_3-C_4 intermediates, cementing the high impact of the minimal change of *GLDP* expression (Rawsthorne et al., 1988b, Morgan et al., 1993).

Despite the photosynthetic gains of the C_2 shuttle, glycine metabolization also releases ammonia into the BS cells, which can disrupt cellular homeostasis if not adequately regulated (Monson & Rawsthorne, 2000). A compelling hypothesis posits that the evolutionary pressure to rectify nitrogen imbalance is what drives the evolution towards C_4 photosynthesis, as many C_4 enzymes are installed to refix ammonia into the form of amino acids and shuttle them back to the mesophyll (Sage et al., 2012; Edwards, 2014; Mallmann et al., 2014; Bräutigam & Gowik, 2016).Three alternative pathways have been identified: (i) a glutamate 2-oxoglutarate shuttle, (ii) an alanine pyruvate shuttle, and (iii) an aspartate malate shuttle (Dal'Molin et al., 2010; Mallmann et al., 2014; Bräutigam & Gowik, 2016). In all three cases, ammonia is fixed into amino acids in BS cells by the corresponding aminotransferases, which also exist in C_3 plants but are recruited for new functions in C_3-C_4 and C_4 leaf tissues (Aubry et al., 2011; Mallmann et al., 2014; Schlüter et al., 2018). The latter two pathways are typical C_4 pathways that also bring CO_2 to the BS, being called a weak C_4 cycle when they occur in complement to the glycine shuttle (Lundgren et al., 2016). Metabolic models predict that strengthening the weak C_4 cycle should directly translate to biomass gains, generating a selection pressure that should eventually replace the glycine shuttle (Heckmann et al., 2013; Mallmann et al., 2014). The final result should then have higher efficiency in water and nitrogen-use than C_3-C_4 photosynthesis

(Ghannoum et al., 2009; Ghannoum et al., 2011), which already reaches efficiency gains in some species (Ueno et al., 2006, Vogan & Sage, 2012, Schlüter et al., 2017).

Metabolic background

The metabolic background of an individual or species refers to the set of chemical compounds, or metabolites, that are present in its cells, being involved in various biochemical processes. The presence and quantity of these metabolites give insight into the existing metabolic pathways and can reveal differences in physiological states or adaptations between species. Though preliminary exploration of the metabolic background of C_3–C_4 was performed on the *Flaveria* genus (Mallmann *et al., 2*014), initial metabolite analysis has already identified differential concentrations of alanine, glycine, GABA, gluconic acid, leucine, malate, malonic acid, and valine between C_3 and C_3-C_4 *Moricandia* species (Schlüter et al., 2017). The primary suspects, glycolate and glycerate (Mallmann et al., 2014), are not especially significant in C_3-C_4 *Moricandia*, and aspartate had significant concentration difference for only one C_3-C_4 species, hinting at either diverse pathways or distinct stages of progression towards C_4 (Schlüter et al., 2017). Questions remain about the extent to which these pathways might be convergent across multiple lineages, and if they all handle nitrogen imbalance between BS and mesophyll the same way. Further investigation is needed to clarify the role of these metabolites in other independent origins of the C_3-C_4 trait.

C_3-C_4 as intermediary stage or as evolutionary endpoint?

C_3-C_4 photosynthesis is much rarer than C_3 or C_4 photosynthesis, and it often occurs in plant lineages where C_4 has also developed (Sage et al., 2018; Lundgren, 2020). Nonetheless, its' existence is probably underestimated as it requires complex gas exchange equipment and analysis, usually brought about in studies of known C_4 (Sage et al., 2011; Lundgren, 2020). Evidence for the convergent evolution of the C_2 shuttle is increasingly clear, with independent examples happening in 11 plant families, both monocots and eudicots, including Brassicaceae, Asteraceae and Poaceae (Schulze et al., 2016; Lundgren, 2020). Furthermore, the variation in stages of progress towards C_4, that can be seen in naturally existing C_3-C_4 species lends to the idea of a stepwise evolutionary path (Heckmann et al., 2013; Blätke & Bräutigam, 2019; Edwards, 2014; Edwards, 2019). Despite this, C_3-C_4 photosynthesis may not presently be excluded as an evolutionary endpoint on its own (Lundgren, 2020; Walsh et al., 2023). Supporting that theory is the fact that some lineages have developed a C_2 carbon shuttle and

remained without transitioning to the C_4 shuttle for over 10 million years (Walsh et al., 2023). Indeed, photorespiration is actually net positive under certain conditions (Timm & Bauwe, 2013; Eisenhut et al., 2019; Broncano et al., 2023), and its biochemical trade-offs may limit the evolution towards C_4 in fluctuating environments (Walsh et al., 2023). The protective role of photorespiration against abiotic stress conditions is known, usually through redox reactions (Wingler et al., 2000; Voss et al., 2013). It can rapidly alleviate oxidative stress generated by high salinity (Mittova et al., 2003; Abogadallah, 2011; Yu et al., 2011), by chilling (Guo et al., 2006; Cheng et al., 2007) or exposure to heavy metals (McCarthy et al., 2001; Romero-Puertas et al., 2002; Cai et al., 2011). Photorespiration also helps dissipating excess energy that could cause photoinhibition in cases of excess light (Bauwe et al., 2010; Eisenhut et al., 2019), or photo-damage during drought stress (Haupt-Herting & Fock, 2002; Guan et al., 2004). Furthermore, there are indications that C_4 photosynthesis loses efficiency under fluctuating light conditions, which happens both due to weather or canopy coverage (Li et al., 2021), where CO_2 fixation rates are more positive for C_3-C_4 and C_3 species (Li et al., 2021). In these specific conditions, C_3-C_4 photosynthesis may strike an ideal balance between positive photorespiration, and reduced oxygenation function of RuBisCO, potentially even fitter than C_4 (Walsh et al., 2023). In sum, C_3-C_4 photosynthesis is relevant not only as a probable intermediary stage that will help us fully understand C_4 evolution, but equally as a standalone improvement on C_3 photosynthetic output, especially advantageous in fluctuating environmental conditions.

Relevance for agriculture

Crop breeding and engineering

Improving photosynthetic efficiency in crops can be achieved with plant breeding (Araus et al., 2021). Several studies have explored ways to improve photosynthetic efficiency in crops, including genetic engineering of key enzymes, and the implementation of more efficient carbon fixation and photorespiration pathways. As a clear illustration, a recent study by South et al. (2019) used gene editing to introduce glycolate metabolism enzymes from *E. coli* and algae into the photorespiratory cycle of tobacco plants, which significantly increased photosynthetic rates and biomass production. The earliest attempts at introducing single C_4 genes in rice failed to improve photosynthetic CO_2 assimilation rates and resulted in stunted morphologies and metabolic energy waste via PPDK, MDH or ME, PEPC enzymes (Taniguchi et al., 2008). Another study by Wang et al. (2017) used maize genes to induce chloroplast and mitochondrial development in rice BS cells, in an attempt to implement a proto-Kranz anatomy (Sage et al.,

2012; Sage et al., 2013), achieving neither fitness gains nor losses in terms of in leaf CO_2 concentration or photosynthetic output. Complicating the picture, individual genes don't always have a linear relation with phenotypes. Liu et al (2023) identified genes via knock-out that promote vein density, noting that only the knock-out of pairs of genes generated an effect, while knocking out each individual one did not. Overall, improving photosynthetic efficiency in crops is a complex and challenging task, but one that has the potential to significantly improve agricultural productivity and sustainability.

Many of the top agricultural crops utilize C_3 photosynthesis, requiring moderate to high amounts of water and nitrogen (Makino, 2011). The lower water and nitrogen requirements of the C4 phenotype are often in focus as relevant for agriculture, as leading to higher yields, resistance to environmental stress and even faster growth (Schuler et al., 2016; Atkinson et al., 2016). Yet the changes required for a total C_4 transformation are complex and not yet fully understood (Sedelnikova et al., 2018). Furthermore, some other commercial crops belong to plant families where C4 seems to not naturally occur, like Brassicaceae, potentially due to anatomical or ecological limitations (Schlüter et al., 2017). C_4 photosynthesis itself is disadvantageous under certain conditions, such as high atmospheric CO_2 content, low light, and cooler temperatures (Bellasio & Farquhar, 2019). On the other hand, implementing a C_2 glycine shuttle can be achieved with a few changes to the regulation of already existing genes (Gowik & Westhoff, 2011; Lundgren, 2020). Nonetheless, the impact of the C_2 glycine shuttle in singlehandedly improving water- and nitrogen-use efficiency is inconsistently observed across taxa: While some C_3-C_4 species exhibit higher net photosynthetic rates than closely related C_3 species (Monson, 1989; Ueno et al., 2006; Voznesenskaya et al., 2007; Vogan et al., 2007; Vogan & Sage, 2011), others do not (Schlüter et al., 2017), though the environmental conditions under which these traits are tested are often limited (Vogan & Sage, 2011). The forming consensus is that C_3-C_4 represents minimal improvements to C_3 in photosynthetic and water-use efficiency under mild temperatures (25°C), but that it offers clear advantages in efficiency and resilience at high temperatures (35°C) or variable conditions (Bellasio & Farquhar, 2019). This would make the transformation of C_3 crops into C_3-C_4 relevant for heat-stressed regions and variable climates, alleviating pressures on global food production systems, reducing water and nitrogen requirements (Mercado & Studer, 2022). Reduced water and nitrogen usage during crop growth would alleviate the pressures on ground and freshwater resources, as well as decrease the ecological footprint of farming practices (Smith et al., 2023). In the face of global food insecurity (Tilman et al., 2011), climate volatility (Malhi et al., 2021), and the need for

sustainable land management (Pérez-Soba et al., 2008), C_3-C_4 photosynthesis represents a step toward a more resilient and less environmentally damaging agricultural paradigm (Smith et al., 2023). Finally, the practicality of transformation into C_3-C_4 or even to C_4 should be considered on a crop-specific basis. For example, the low leaf area-to-ground area ratio of crops like rice diminishes the benefits of C_4 photosynthesis due to shading (Bellasio & Farquhar, 2019). Ultimately, though the development of a full C_4 cycle may be the most robust solution for some crops in consistently hot and arid climates, C_3-C_4 also offers potential improvements in crop resilience to abiotic stress and productivity, as well as being a bridge for future C_4 development (Mercado & Studer, 2022).

Genomes and mechanisms of genetic regulation

The level of understanding of C_4 and C_3-C_4 photosynthesis has reached the frontier of DNA and RNA. Unraveling genetic information encoded within plant genomes has become a cornerstone of biological research, enabling a deeper understanding of plant development, adaptation, and evolution (Michael & Jackson, 2013).

Gene regulatory elements and transcription factors

In plants, as well as in animals, the differentiation of cell types from the same genome is directed by many concurrent mechanisms of genetic regulation (de-Leon & Davidson, 2007). Coordination is mainly driven by regulatory proteins latching onto promoters, enhancers, and other regulatory elements in the genome (Shlyueva et al., 2014). These regulatory elements are small DNA segments in the genome that bind with transcription factors (TFs), which in turn recruit cofactors that locally modify chromatin structure and kickstart or block gene transcription (Shlyueva et al., 2014). A gene is usually influenced by several regulatory elements, coded in many possible relative positions in the same (cis) DNA strand (de-Leon & Davidson, 2007; Long et al., 2016). The variety in regulatory elements opens multiple pathways for divergent evolution in the expression of a gene, underpinning variety between individuals or species.

Importance of non-coding genomic variation for gene expression

Another source of variation in gene expression has to do with the chromosome unfolding during transcription to RNA (Vergara & Gutierrez, 2017; Bernardi, 2021). What was in the past called

junk DNA, the large part of the genome that does not code for genes, is increasingly seen as vital for the gene regulation machinery: the distinctive chemical properties of CG and AT mean that the amino-acid composition of DNA affects how it folds, rendering specific regions more accessible to transcription factors (Bernardi, 2021). Furthermore, epigenetics mechanisms like demethylation and methylation of CpG islands are respectively associated with active gene transcription or gene silencing (Deaton & Bird, 2011). Such mechanisms for example play a crucial role in the expression of photosynthesis-related genes like *RuBisCO* and PEPC, acting with in response to environmental factors like light and drought (Duarte-Aké et al., 2019). This illustrates the vital role of non-coding regions in shaping gene expression and regulatory processes across different cell types.

Transposons as sources of genomic variation

There is a plethora of ways wherein genomic evolution can affect gene expression, from the direct creation or inactivation of TF recognition sites to rearrangement of the chromatin structure and unfolding (Mardis, 2007; Hirsch & Springer, 2017). In plants, one of the biggest sources of genomic structural variation are transposons (Hirsch & Springer, 2017). These are DNA segments which move within the genome across generations, by virtue of recognition sites for enzymes that mediate their movement. DNA transposons bind directly to Transposase enzymes, which cut-and-paste them somewhere on the genome, while retrotransposons first get transcribed and then bind to reverse transcriptases, which write a DNA copy of the transposon on the genome (Bourque et al., 2018). The dynamic nature of transposons introduces diversity into genomes, occasionally resulting in alterations in gene expression and, consequently, distinct phenotypes between individuals or across various cell types (Hirsch & Springer, 2017). Not only can transposon mobility lead to the activation or inactivation of single genes, but it may also extend beyond single genes to affect the regulatory networks governing cellular functions (Bourque et al., 2018). The intricate interplay between transposons and the genome wields a transformative influence on genomic evolutionary processes. This can be seen in C_3-C_4 evolution, where spatially differentiated gene expression patterns often stem from transposons. For example, the rearrangement of upstream regulatory regions by transposons could contribute to the differential expression of genes involved in glycine shuttling or decarboxylation pathways (Hirsch & Springer, 2017). Transposons are also thought to play an evolutionary role in recruiting enzymes and regulatory elements to the C_4 cycle (Cao et al., 2016; Zhu et al., 2024). These processes illustrate how genomic plasticity drives the evolution of complex traits like C_3-C_4 and C_4 photosynthesis, emphasizing the need for studying the

genome as a whole and not just separate genes, which depends on the complex process of genome sequencing and assembly.

Foundations for genomics research

In recent decades, the field of plant genomics has undergone a remarkable transformation, primarily driven by advancements in sequencing technologies. Emergent high-throughput sequencing platforms revolutionized the field, facilitating the rapid and cost-effective analysis of plant genomes and transcriptomes with unprecedented accuracy and resolution (Kress et al., 2022). Sequencing technologies have diverged to accommodate different research needs. Short read sequencing is a cost effective method for high throughput whole genome sequencing (Quail et al., 2008) but struggles to resolve repetitive or highly heterozygous regions of genomes, common in plants (Treangen & Salzberg, 2012). Long read and linked read sequencing are both excelling solutions for contiguous genome assembly, full-length transcriptome analysis, and haplotype phasing (Mostovoy et al., 2016; Pollard et al., 2018). Proximity ligation sequencing methods, such as Hi-C, reflect the genome three-dimensional architecture (Lieberman-Aiden et al., 2009), useful both for scaffolding assemblies into chromosome level contiguity and for unveiling higher-order chromatin structures like loops and topologically associated domains (TADs) (Beagan et al., 2020). Hi-C has proven instrumental in annotating regulatory elements, deciphering long-range interactions, and unraveling the intricate folding dynamics of genomes. Finally, when having quality reference genome assemblies, higher level of detail, ChIP-seq and Hi-ChIP represent pivotal techniques for dissecting protein-DNA interactions, instrumental in uncovering transcription factor binding sites, histone modification patterns, and other chromatin-associated proteins, which provides critical insights into the regulatory networks governing gene expression (Mardis, 2007). These advancements in genomic technologies collectively contribute to a more comprehensive understanding of the functional and structural aspects of the genome, shedding light on the regulatory mechanisms that orchestrate cellular processes.

Most levels of analysis require knowing gene locations in the genome. Several methods are established for predicting genes *ab initio* out of genomic sequences, using Hidden-Markov-Models and optionally mapped RNA transcripts, which infer DNA patterns associated with genes (Stanke & Waack, 2003; Johnson et al., 2008). Moreover, we currently see the dawn of Deep Learning models for easier and more accurate gene annotation (Jaganathan et al., 2019; Wang et al., 2019; Stiehler et al., 2020). The field of transcriptomics in particular has allowed for mass identification of genes, with clues on their function based on differential gene

expression between samples of different species, conditions, tissues or cell types (Weber, 2015). In this way, many genes relevant for C_3-C_4, C_3 and C_4 photosynthesis have already been identified (Bräutigam et al., 2011; Stata et al., 2016; Lauterbach et al., 2017). However, the picture remains incomplete without another layer of knowledge: How is differential expression achieved across cell types? How is gene expression changed in response to environmental factors? What polymorphisms map to phenotypic differences across individuals? These are questions that can only be inspected with the help of genome assemblies.

A growing repertoire of C_3-C_4 and C_4 genomes and transcriptomes is asserting itself, with examples in the families Poaceae (Emrich et al., 2004; McCormick et al., 2018), Asteraceae (Wang et al., 2013; Bonthala et al., in preparation) and Amaranthaceae (Wang et al., 2019). With genome assemblies and accurate gene annotations, the groundwork for more complex studies is laid, allowing researchers to study gene spatial arrangements, alternative splicing, structural and polymorphism variations, promoter regions and repetitive elements. Nuanced differences in any of these features can have effects in disease resistance, stress tolerance, cell differentiation and developmental processes (Gabur et al., 2019). Future transcriptomics studies are reinforced by the existence of reference genomes, as differential expression analysis is affected by read depth estimations, which are best when mapped against a genome (Burgess & Hibberd, 2015). Sequenced genomes also allow the design of molecular markers used to probe variation across different individuals (Zhu et al., 2023), based on the single nucleotide polymorphisms (SNPs) (Ganal et al., 2009, 2012). It is now possible and relatively cheap to genotype plant individuals by sequencing and screening with SNP arrays (Clarke et al., 2016). Making use of population statistics, Genome Wide Association Studies (GWAS) and Quantitative Trait Loci (QTL) mapping are valuable tools to pinpoint genomic areas of interest related to a phenotype in question (Yano et al., 2016; Clarke et al., 2016). Through these magnifying lens, regulatory mechanisms, or even elusive novel genes (Yano et al., 2016) may be uncovered. Finally, multiple genome comparison allows the identification of conserved regions and evolutionary relationships across different plant breeds or species. Such investigations have provided valuable insights into the evolutionary history of plants, highlighting shared genetic features and evolutionary adaptations.

Phylogenetics of C_3-C_4 photosynthesis

Phylogenetics provides a robust foundation for the formulation and empirical testing of scientific hypotheses, as it quantifies the relatedness between species. This relatedness is contingent on the species temporal separation, which is followed by genomic divergence both

due to distinct evolutionary pathways and due to accumulation of different random mutations (Brocchieri, 2001). Knowing species' relatedness allows an informed comparison of their genomes. Conserved genomic regions observed across distantly related species often signify segments of biological significance that are critical to the viability and fitness of organisms (Bejerano et al., 2005). Conversely, regions of high genetic divergence between closely related species may signify areas that have been subjected to distinct evolutionary pressures and may correlate with phenotypic divergence (Massingham & Goldman, 2005). It is however imperative to exercise caution when interpreting such correlations, as they easily yield false positives. Comparing several species at the same time and accounting for their evolutionary relatedness increases statistical power and is essential to discern whether divergences may be biologically meaningful or merely incidental artifacts (Nagy et al., 2020). This must be followed by functional scrutiny of the DNA segments and what they code for.

Much work has been done on C_4 phylogenetics, with estimations of over 40 independent evolutions of the trait (Sage, 2004; Sage et al., 2011) since it first appeared 25–30 million years ago as an adaptation to an atmospheric decrease in CO_2 (Christin et al., 2008; Christin et al. 2011). C_3-C_4 photosynthesis is starting to attract more attention, included in small phylogenies encompassing *Flaveria* (McKown et al., 2005; Munekage & Taniguchi, 2022) or *Moricandia* (Schlüter et al., 2017) and Poaceae (Grass Phylogeny Working Group II., 2012; Khoshravesh et al., 2016). Lundgren et al. (2017) performed an angiosperm-wide joint analysis of C_3, C_3–C_4, and C_4 species, to test the relation between photosynthesis kind and ecological niche. Controlling for phylogenetical effects, they concluded that C_3-C_4 taxa tend to inhabit warm climates with more seasonal rainfall, independent of their ancestral condition. They also noted that almost all C_3-C_4 species have close C_4 relatives, apart from species in the Brassicaceae family. They supposed that the geographical distributions of this family in areas with relatively high minimum precipitation and fertile soil and relatively colder temperatures decrease the evolutionary pressure towards full C_4 evolution. Further investigations in a Brassicaceae context could yield practical insights for commercial applications, as well as could confirm or disprove the hypothesis that the C_2 carbon pump evolutionary endpoint on its own.

C_3-C_4 in Brassicaceae

Brassicaceae evolved ~23 million years ago around the Mediterranean (Arias & Pires, 2012) and hosts C_3-C_4 species in at least two genera (Apel et al., 1997). Already in the infancy of the development of Carbon Compensation Point (CCP) measurements, more than 60 years ago

(Hatch et al., 1967), it was noted that *M. arvensis* did not fit in either a C_3 or a C_4 pattern (Krenzer, 1975). The Brassiceae tribe, one of the largest and most commercially important tribes within Brassicaceae, includes staple crops such as cabbage, broccoli, rapeseed, turnip and mustard. Being the family of *Arabidopsis thaliana* and many commercial vegetables, Brassicaceae has already been subject of much phylogenetical work and discussion (Warwick & Black, 1991; Al-Shehbaz et al., 2006; Arias et al., 2014; Nikolov et al., 2019). However, such a speciose family is hard to fully cover, and many of the C_3-C_4 species are wild weeds that are ignored in favour of more charismatic species. Some C_3-C_4 species are poorly known and are often ignored in holistic studies and phylogenies, like *B. gravinae* (Ueno, 2011) and *D. erucoides*, whereas most C_3-C_4 focus is on the *Moricandia* and *Diplotaxis* genera that have very close C_3 relatives (Rawsthorne et al., 1988a, 1988b; Schlüter et al., 2017; Pinheiro et al., 2023). Phylogenetical work and systematic phenotype characterization are required for a better inspection of C_3-C_4 in this family. The most complete work done before this thesis was performed by Schlüter at al. (2017). It analysed *Moricandia* in detail, suggesting a single origin of the glycine shuttle in an ancestor of the 3 known intermediate species. However, the phylogeny was based only on nuclear ITS markers, which have accuracy limitations when compared to a more comprehensive approach involving multiple genes (Degnan & Rosenberg, 2009). Previously it was suggested that the genetic mechanisms governing the C_3-C_4 anatomy and physiology of *Moricandia* are different than those existing in C_4 (Rylott et al. 1998). It would be interesting to see if that extends to the other C_3-C_4 species in Brassicaceae.

Being the family of veteran model species is a great advantage for the study of more complex patterns of gene expression (Mabry et al., 2023). Not only does *A. thaliana* possess gold standard annotations for genes and understudied classes of non-coding RNA like smallRNA, microRNA, long intergenic RNA, small nucleolar RNA, natural antisense transcripts and small nuclear RNA (Cheng et al., 2017), it has also been the subject of extensive exploration of gene enhancers (Meng et al., 2021) and of connection between DNA methylation and gene expression (Zicola et al., 2019; Zhao et al., 2020). Specific traits connected with C_3-C_4 photosynthesis have already been approached: For example, in *A. thaliana*, the genetic mechanisms controlling vein density are relatively well known (Rishmawi et al., 2017) and epigenetic mechanisms modulating BS cell anatomy and chloroplast content have recently been identified (van Rooijen, 2020). The promoter region of *GLDP1* has already been scrutinized in this species, with the identification of an M-box regulatory element that might be related with bundle sheath expression (Adwy et al., 2015, 2019). Finally, *A. thaliana* is subject of accumulated experience with genome editing (Meng et al., 2021; Miki et al., 2021) and

population statistics (Rishmawi et al., 2017), valuable tools in genomic analysis.

Aims of Thesis

As established, the Brassicaceae family and particularly its Brassiceae tribe, is a relevant and underappreciated subject for C_3-C_4 research. Conclusions on how to instantiate C_3-C_4 evolution in this family would generate agricultural gains at a fraction of the difficulty of engineering a full C_4 cycle in these plants (Lundgren et al., 2020). To comprehensively understand C_3-C_4 photosynthesis, a multifaceted approach encompassing both physiological and genomic aspects is essential. Therefore, this thesis aims for the following:

- Establish and annotate a comprehensive panel of Brassicaceae genomes, covering both C_3-C_4 species and their close C_3 relatives, facilitating a robust comparative analysis.
- Perform detailed quantitative description of C_3-C_4 characteristics across the panel, including Carbon Compensation Point (CCP) and related anatomical and physiological traits.
- Employ phylogenetic methods to elucidate the evolutionary history of C_3-C_4 intermediacy in Brassicaceae.
- Describe metabolic background and its variability to identify key metabolites related to the C_3-C_4 trait.
- Investigate genomic regulation mechanisms influencing the C_3-C_4 differentiation of mesophyll and bundle sheath (BS) cells.
- Explore the practical implications of C_3-C_4 photosynthesis in Brassicaceae for crop improvement, with an emphasis on the potential for genetic engineering.

Publications

1. A genomic panel for studying C_3-C_4 intermediate photosynthesis in the Brassiceae tribe

Except for Figure 6, all bioinformatical work here performed was by me under the supervision of Prof. Stich and Dr. Bonthala. Laboratory work was mainly performed by laboratory technicians Stephanie Krey and Agata Stoltmann. Some of the plant specimens were procured and grown by me with the help of Stephanie Krey, while others were grown previously by Agata Stoltmann. Library preparation and sequencing were performed by 3rd parties, whereby I organized the shipping logistics and sequencing process, with the help of Prof. Stich and Stephanie Krey. The text was mainly written by me. Collaborators helped with ideas, as well as providing physiological and sequencing data for some species, and with manuscript reviewing, especially in the introduction.

Received: 2 March 2023 | Revised: 18 May 2023 | Accepted: 23 June 2023

DOI: 10.1111/pce.14662

DATA INSIGHT

Plant, Cell & Environment | WILEY

A genomic panel for studying C3-C4 intermediate photosynthesis in the Brassiceae tribe

Ricardo Guerreiro[1] | Venkata Suresh Bonthala[1] | Urte Schlüter[2,3] | Nam V. Hoang[4] | Sebastian Triesch[2,3] | M. Eric Schranz[4] | Andreas P. M. Weber[2,3] | Benjamin Stich[1,3,5]

[1]Institute of Quantitative Genetics and Genomics of Plants, Faculty of Mathematics and Natural Sciences, Heinrich Heine University, Düsseldorf, Germany

[2]Institute of Plant Biochemistry, Faculty of Mathematics and Natural Sciences, Heinrich Heine University, Düsseldorf, Germany

[3]Cluster of Excellence on Plant Sciences (CEPLAS), Düsseldorf, Germany

[4]Biosystematics Group, Department of Plant Sciences, Wageningen University, Wageningen, The Netherlands

[5]Max Planck Institute for Plant Breeding Research, Köln, Germany

Correspondence
Benjamin Stich, Institute of Quantitative Genetics and Genomics of Plants, Faculty of Mathematics and Natural Sciences, Heinrich Heine University, Düsseldorf, Germany.
Email: benjamin.stich@hhu.de

Funding information
Collaborative Research Centre/Transregio (TRR 341), Grant/Award Number: 456082119; Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) in the frame of the ERA-CAPS project C4BREED; Germany's Excellence Strategy (EXC 2048/1), Grant/Award Number: 390686111

Abstract

Research on C4 and C3-C4 photosynthesis has attracted significant attention because the understanding of the genetic underpinnings of these traits will support the introduction of its characteristics into commercially relevant crop species. We used a panel of 19 taxa of 18 Brassiceae species with different photosynthesis characteristics (C3 and C3-C4) with the following objectives: (i) create draft genome assemblies and annotations, (ii) quantify orthology levels using synteny maps between all pairs of taxa, (iii) describe the phylogenetic relatedness across all the species, and (iv) track the evolution of C3-C4 intermediate photosynthesis in the Brassiceae tribe. Our results indicate that the draft de novo genome assemblies are of high quality and cover at least 90% of the gene space. Therewith we more than doubled the sampling depth of genomes of the Brassiceae tribe that comprises commercially important as well as biologically interesting species. The gene annotation generated high-quality gene models, and for most genes extensive upstream sequences are available for all taxa, yielding potential to explore variants in regulatory sequences. The genome-based phylogenetic tree of the Brassiceae contained two main clades and indicated that the C3-C4 intermediate photosynthesis has evolved five times independently. Furthermore, our study provides the first genomic support of the hypothesis that *Diplotaxis muralis* is a natural hybrid of *D. tenuifolia* and *D. viminea*. Altogether, the de novo genome assemblies and the annotations reported in this study are a valuable resource for research on the evolution of C3-C4 intermediate photosynthesis.

KEYWORDS
Brassicaceae, C3-C4 intermediate photosynthesis, evolution

1 | INTRODUCTION

Carbon concentrating mechanisms enable plants to reduce photorespiration and improve their photosynthetic efficiency especially under conditions of high temperatures and limited water supply (Bellasio & Farquhar, 2019; Sage et al., 2012). In C4 photosynthesis, a high CO_2 atmosphere is achieved in the bundle sheath cells by complex modifications of leaf biochemistry, anatomy and ultrastructure (Hatch, 1987). C4 photosynthesis is therefore not only the focus of fundamental research but also crop breeding programmes may benefit from a better knowledge of the trait (Schuler et al., 2016). However, our understanding of the genetics underlying C4 photosynthesis is still very fragmented and attempts to introduce C4 traits into agriculturally relevant crop species were only partially successful (Ermakova et al., 2021; Wang et al., 2017). An alternative approach might, therefore, focus on the understanding of carbon concentration through the glycine shuttle mechanism, a pathway that is supposed to represent an early step during the evolution from C3 to C4 photosynthesis (Mallmann et al., 2014; Rawsthorne et al., 1992). Plants employing the glycine shuttle mechanism are often termed C3-C4 intermediates or C2 species because a C2 compound is exchanged between the cells (Edwards & Ku, 1987; Rawsthorne et al., 1992; Sage et al., 2014). Measurable parameters such as the CO_2 compensation point or mitochondria and chloroplast accumulation in bundle sheath cells usually show intermediate values between C3 and C4 plants (Ku et al., 1991; McKown et al., 2005; Muhaidat et al., 2011). Biochemical, anatomical and ultrastructural modifications in the C3-C4 leaf are therefore likely to be less complex than in C4 plants, easier to understand and, thus, to engineer (Bellasio & Farquhar, 2019; Lundgren, 2020).

The photorespiratory cycle describes the recycling of 2-phosphoglycolate (2PG), a toxic metabolite that is formed when Rubisco reacts with oxygen instead of CO_2. 2PG is initially converted into glycolate in the plastids and transported into the peroxisome. There it is further metabolised into glyoxylate and aminated into glycine. In the mitochondria, two molecules of glycine are converted by coordinated reactions of the glycine decarboxylase complex and the serine hydroxymethyl transferase into one molecule of serine, CO_2 and NH_3 (for recent reviews see: Eisenhut et al., 2019; Timm and Hagemann, 2020). Through further reactions taking place in the peroxisome and plastid, serine is deaminated into hydroxypyruvate, then metabolised into glycerate and finally converted into 3-phosphoglycerate, a metabolite that can enter into the Calvin-Benson-Bassham cycle. In C3 species, the complete photorespiratory cycle takes place in all photosynthetically active cells of the leaf. Shifting the glycine decarboxylation step exclusively to the bundle sheath cells leads to increased CO_2 release in these cells, creating an elevated CO_2 environment where the oxygenase reaction of Rubisco is considerably reduced. The bundle sheath specific localisation of the P-protein from the glycine decarboxylase complex has been shown in C3-C4 species from diverse phylogenetic backgrounds by immunolocalization (Khoshravesh et al., 2016; Oono et al., 2022; Rawsthorne et al., 1988; Schlüter & Weber, 2016). The glycine shuttle biochemistry is accompanied by enhanced centripetal organelle accumulation in the bundle sheath cells (for reviews see: Lundgren, 2020; Schlüter and Weber, 2016). Carbon concentration via the glycine shuttle is less effective than the C4 cycle, but could be advantageous under hot and dry growth conditions, when photorespiration is usually high (Bellasio & Farquhar, 2019; Oono et al., 2022; Schlüter et al., 2023; Vogan & Sage, 2012). Since the C3-C4 related features could represent transitory stages towards C4 photosynthesis, knowledge of their genetic underpinnings could also contribute to the understanding of C4 evolution.

The anatomical and physiological differences between C3 and C3-C4 intermediate species are relatively well studied and characterised in the Brassicaceae genus Moricandia (Schlüter et al., 2017). Genetic factors responsible for these differences have mostly been analysed through the lens of transcriptomics (Bräutigam et al., 2011; Gowik et al., 2011; Lauterbach et al., 2017; Schlüter et al., 2017, Siadjeu et al., 2021). While transcriptome analysis unravels gene expression patterns, it alone is not sufficient for understanding gene regulatory mechanisms (Conant et al., 2014). Therefore, as an extra layer of information, whole genome assemblies can be used with comparative and quantitative approaches to investigate the regulatory genes and elements, genome duplications and structural variations (Adwy et al., 2015, 2019; Conant et al., 2014, Schulze et al., 2013). The existence of genome assemblies also facilitates other classical and modern methods for genetic inspection and analysis, genome editing (Jeong et al., 2019), resequencing, gene expression assessment, as well as genetic mapping of phenotypic variation.

Most effort has so far been put into understanding C4 photosynthesis in phylogenetically disparate species such as maize (Denton et al., 2017; Wang et al., 2013), *Gynandropsis gynandra* (Külahoglu et al., 2014; Reeves et al., 2018) or *Flaveria* sp. (Gowik et al., 2011; Taniguchi et al., 2021) and implementing the complete C4 trait into quite distantly related but agriculturally relevant C3 species such as rice (von Caemmerer et al., 2012; Ermakova et al., 2021; Schulze et al., 2016). Understanding how to convert C3 into C3-C4 photosynthesis is less challenging and could already produce commercially relevant yield gains (Lundgren, 2020; Schulze et al., 2016; Weber & Bar-Even, 2019). In this context, the Brassicaceae family is intriguing as it contains the genetically very well characterised model species *Arabidopsis thaliana* and commercially relevant species such as *Brassica napus* (canola) and *B. oleracea* (cabbage). In addition, this family also includes multiple C3-C4 intermediate evolutionary lineages (Apel et al., 1997; Sage et al., 2011). Hence, Brassicaceae species are ideal for investigating C3-C4 evolution in a pan-genomic context to understand the differences in gene regulation and studying convergent evolution. In addition, Brassicaceae species are known to produce fertile progenies in interspecific crosses (Kaneko & Bang, 2014; Ueno et al., 2003). Such progenies can be helpful for unravelling the inheritance of C3-C4 intermediacy and for transferring the genes of interest to relevant crops.

The main aim of this study is to establish the genomic resources that enable comparative genetic and genomic research on C3-C4 intermediate photosynthesis. In detail, the objectives of our study were to:

(1) create draft genome assemblies and annotations of 19 closely related Brassiceae taxa with different photosynthesis characteristics (C3 and C3-C4),
(2) quantify orthology levels using synteny maps between all pairs of taxa,
(3) describe phylogenetic relatedness across all the taxa, and
(4) track the evolution of C3-C4 intermediate photosynthesis in the Brassicaceae family.

2 | MATERIALS AND METHODS

2.1 | Genetic material

We collected seeds for 18 Brassiceae species (19 taxa) from various gene banks (Table S1), for which the genome sequences were unavailable. Thereby we considerably increased the coverage of this tribe in which C3-C4 intermediacy has been reported previously. A subset of the above mentioned taxa was selfed one to several times to reduce heterozygosity and facilitate genome assembly.

2.2 | Linked-read library preparation and sequencing

For all 19 taxa, linked-read sequencing was performed using either 10x (Zheng et al., 2016) or stLFR (Wang et al., 2019) technologies (Table S1). Initially, for 15 taxa (Table S1), DNA was extracted with the DNeasy Plant Mini Kit (QIAGEN) following the manufacturer's instructions and size-selected for fragments larger than 40 Kb using BluePippin (SAGE Sciences). Quality control of the size-selected DNA was performed on Qubit and TapeStation. A 10x linked-read library (Zheng et al., 2016) was created for each taxa using 1 ng of DNA as recommended by the manufacturer. Sequencing was performed on the HiSeq. 3000 sequencer with pair-end mode by Novogene.

For the remaining four taxa (Table S1), stLFR linked-read libraries (Wang et al., 2019) were prepared by BGI from tissue samples using MGIEasy stLFR Library Prep Kit (MGI). The libraries were sequenced on BGISEQ-500 (100 bp and pair-end) by BGI. In addition, we re-sequenced one species due to unsatisfactory quality of 10x data, using stLFR link-read technology as mentioned above (Table S2).

2.3 | Long-read library preparation and sequencing

Complementary long-read data was generated for a subset of seven taxa (Table S1) to improve the de novo genome assemblies. PacBio SMRTbell libraries were prepared as recommended by Pacific Biosciences (SMRTbell Template Prep Kit 1.0 SPv3), including a size selection on Blue Pippin to remove fragments lower than 10 Kb. Sequencing was performed on Sequel with 2.0 Binding Kit and sequencing chemistry for 10 h, or 3.0 Binding Kit and sequencing chemistry for 20 h, as recommended by Pacific Biosciences. Oxford Nanopore libraries (Table S1) were prepared from purified high molecular weight DNA extracted from leaf tissue by precipitation of DNA-CTAB complexes (Arseneau et al., 2017; Xin & Chen, 2012). In a second step, CTAB was removed with ethanol, and the co-purified RNA was digested by RNAse treatment. Afterwards, the DNA was again purified by binding it to AMpure PB (Pacific Biosciences) beads, washing the beads in ethanol and then resolving the DNA. Sequencing was performed by GridION and PromethION flow cells by GTL Düsseldorf.

2.4 | Estimation of genome size, heterozygosity and repeat content

From the linked-read libraries of each taxa, the 21-mers were extracted using Jellyfish (version 2.1.3) (Marçais & Kingsford, 2011). Genomescope (www.genomescope.com) was then used to estimate genome size, heterozygosity and repeat content, setting the maximal k-mer coverage parameter to 10 000.

2.5 | Genome assembly

The supernova assembler v2.1.1 (Weisenfeld et al., 2017) was used to assemble both 10x and stLFR linked-read data to pseudohaploid assemblies. Long Ranger v2.2.2 (Ott et al., 2018) was used with default parameter settings to map the linked reads to respective de novo genome assemblies. Purge Haplotigs v1.1.0 (Roach et al., 2018) was used to reduce the under-collapsed haplotigs in all de novo genome assemblies. Deduplication was not possible for *Diplotaxis muralis* due to Long Ranger failing to map the linked reads. BUSCO v3.1.0 (Simão et al., 2015) was used based on the eudicot_db10 database to estimate the completeness of the de novo genome assemblies before and after reducing the under-collapsed haplotigs.

PacBio long reads for *Eruca sativa* and *D. erucoides* were assembled with Canu v1.8 (Koren et al., 2017) using default parameters except for corOutcoverage = 200 and correctedErrorRate = 0.15 and discarding reads shorter than 1000 bp. To deal with the higher error frequency of long reads, the Canu assemblies were polished using Pilon v1.22 (Walker et al., 2014) with the less error-prone linked reads by mapping in two iterations. The polished and purged PacBio assemblies were further scaffolded with the LINKS v1.8.7—ARCS v1.1.1 pipeline (Warren et al., 2015; Yeo et al., 2018) that performs misassembly correction with Tigmint v1.1.2 using linked reads (Jackman et al., 2018).

Oxford Nanopore long reads data obtained for *D. acris*, *D. harra*, *Hirschfeldia incana* HIR3, *Moricandida sinaica* and *M. spinosa* were basecalled with Guppy v5.0.11 (Wick et al., 2019). The resulting reads

were then trimmed on the first 50 bp and filtered with NanoFilt v2.6.0 (De Coster et al., 2018) on a minimum length of 1000 bp and minimum average phred-64 quality score of 10. The high-quality reads were subsequently used for scaffolding linked-read assemblies with LINKS v1.8.7 (Warren et al., 2015) as well as gap filling with NanoFilt v2.6.0. To make sure that sequencing errors were not incorporated to the assemblies, the resulting sequences were afterwards polished with Pilon 1.22, using the linked reads.

Assemblies for *G. gynandra* (Hoang et al., 2023), *M. arvensis* and *M. moricandioides* (Lin et al., 2021) were obtained directly from collaborators. The Moricandia assemblies were also polished with the linked reads from our study. Additional assemblies were available from NCBI (Table S2).

Assembly statistics such as N50-90, L50-90, assembly size and contig number were calculated with a custom python script for each finalised genome assembly.

2.6 | Ploidy estimation

We used nQuire (retrieved in December 2022) (Weiß et al., 2018) to estimate the ploidy of our genomes by analyzing the frequency distribution of biallelic variant sites of reads mapping to BUSCO genes. We generated a histogram of read mapping depths and applied nQuire's denoise tool, which uses a Gaussian Mixture Model (GMM) with uniform noise component approach to remove a uniform baseline from the histogram. The variations in read mapping depth were then used with a GMM to generate a log-likelihood under diploid, triploid, tetraploid and free models. The smallest of the delta-log-likelihoods between the free model and the fixed models was taken as the most likely ploidy.

2.7 | Transcriptome assemblies

RNA-Seq data for *E. sativa* (SRR6454139), *H. incana* (SRR11638396), *D. tenuifolia* (PRJNA904765) and *D. viminea* (PRJNA904804) were downloaded from the Sequence Read Archive database at NCBI, while for *M. arvensis* and *M. moricandioides* were obtained from Schlüter et al. (2017). Trimmomatic v0.39 (Bolger et al., 2014) was used to trim adapters and low-quality reads. Additionally, reads shorter than 36 bp were discarded. The high-quality RNA-Seq reads were then assembled with Trinity v2.11.0 (Haas et al., 2013).

2.8 | Repeat annotation

We performed de novo repeat identification using Marker-P guidelines (Campbell et al., 2014). Briefly, Mite-hunter (Han & Wessler, 2010) with the default parameters were used to identify Miniature inverted-repeat TEs and LTRharvest v1.5.9 (Ellinghaus et al., 2008) with the default parameters for de novo predictions of LTR (Long Terminal Repeat) retrotransposons. Finally, RepeatModeller v1.0.11 (Smit & Hubley, 2015) with default parameters was used to build a de novo repeat library and RepeatMasker v4.0.9 (Smit, Hubley & Green, 2015) was used to mask identified repeats in respective genome assemblies. In addition, repeat annotation was also performed for 13 publicly available species (Table S2).

2.9 | Gene structural annotation

We used protein sequences of all Brassicaceae species available from the UniProt database (The UniProt Consortium, 2019), excluding protein sequences with low evidence levels (Uncertain and Predicted). We searched the UniProt database on 02/10/2020 using the following parameters: taxonomy: Brassicaceae NOT existence: "Uncertain [5]" NOT existence: "Predicted [4]" OR reviewed: yes). We reduced the sequence identity to 95% between any protein sequences present in the downloaded dataset using CD-HIT v4.8.1 (Fu et al., 2012; Li & Godzik, 2006), and the resulting 114 295 protein sequences were used in following gene structural annotation.

Gene structural annotation was performed with Maker2 v2.31.8 (Campbell et al., 2014) in two steps. First, potential genes were annotated based on alignments with the protein sequences in our protein database and the transcript sequences assembled for individual species. Second, the annotated genes were fed to SNAP v2006-07-28 (Korf, 2004) and Augustus v3.3.2 (Hoff & Stanke, 2019) to predict gene structure across all taxa. The model training was performed with Nextflow-abInitio v0.2, made available by National Bioinformatics Infrastructure Sweden, and the trained models were provided in the second run of Maker2.

We initially annotated six taxa: *D. tenuifolia* (C3-C4), *D. viminea* (C3), *H. incana* HIR1 (C3), *M. arvensis* (C3-C4) and *M. moricandioides* (C3) using publicly available RNA transcripts (Mabry et al., 2020; Schlüter et al., 2017) in addition to the protein database mentioned above. The resulting predicted proteins were filtered for AED values smaller than 0.5 and a length >49 amino acids (the minor 1st-percentile protein lengths in *Arabidopsis*). The resulting protein sequences were added to the above-created protein sequence dataset, followed by reducing the sequence identity to 95% using CD-HIT v4.8.1 (Fu et al., 2012; Li & Godzik, 2006). This final protein sequence dataset contained a total of 284 999 proteins. It was used as the only evidence to systematically annotate all genomes of this study with Maker2, including species with existing annotations and the six species we initially annotated. This systematic annotation was done to avoid bias in the downstream analyses with different annotation qualities (Trachana et al., 2011).

2.10 | Gene functional annotation

Functional annotation for the predicted proteins was performed using the Automated Assignment of Human Readable Descriptions (AHRD) v3.3.3 (https://github.com/groupschoof/AHRD). The AHRD

pipeline assigns gene descriptions, Pfam domains (El-Gebali et al., 2019) and Gene Ontology annotations (Barrell et al., 2009; Lewis, 2005) for each gene based on InterProScan v5.42-78.0 (Zdobnov & Apweiler, 2001) and BLASTp v2.9.0+ (Altschul et al., 1990) searches. The BLASTp searches were performed against the Araport11 (Cheng et al., 2017), Swiss-Prot (The UniProt Consortium, 2019) and trembl_plants (O'Donovan, C. et al., 2002) databases (downloaded in 02/2019).

Transposable element (TE) related genes were identified and excluded from the gene annotation (cf. Jayakodi et al., 2020). A predicted protein was labelled as TE-related and filtered out of the protein sequences if at least two out of three fields of the AHRD output, such as AHRD descriptions, the best-blast-hit description or Pfam annotation, were associated with TEs. A list of the terms used for filtering is provided in Table S4.

2.11 | Orthology map and species tree

The TE-filtered protein sequences were analysed by Orthofinder v2.5.1 (Emms & Kelly, 2015, 2019) for orthology identification. Multiple sequence alignments for identified hierarchical orthogroups (HOGs) were produced with MAFFT v7.471 and used for creating gene trees with RAxML v8.2.10 with the PROTGAMMALG substitution model. The gene trees of all HOGs were fed to ASTRAL-pro with default parameters (Zhang et al., 2020) for generating a multi-species coalescent-approach-based species tree.

2.12 | Phylogenetic analysis of *H. incana* accessions based on chloroplast sequences

To ascertain the phylogenetic placement of the two *H. incana* accessions HIR1 and HIR3, in relation to other species in the Brassiceae tribe and also to another accession Nijmegen (NIJ, Garassino et al., 2022), we first assembled their chloroplast genomes using whole-genome sequencing data from our study. Raw reads were trimmed for adapter contamination, quality and length using Trimmomatic v0.39 (Bolger et al., 2014) with the following parameters: "ILLUMINACLIP: 2:20:10 SLIDINGWINDOW:4:15 LEADING:5 TRAILING:5 MINLEN:50". Trimmed reads were used to assemble chloroplast genomes employing GetOrganelle package v1.7.7.0 (Jin et al., 2020) with default setting and the embplant_ptdatabase. For other 12 species, the available chloroplast genomes were downloaded (Table S3). Because chloroplast genomes are still missing for many Brassiceae species, to obtain a higher phylogenetic resolution, we also utilised sequences derived from four rapidly evolving chloroplast intergenic regions, rpl32-trnL, atpI-atpH, psbD-trnT and ycf6-psbM (Arias & Pires, 2012).

To construct the maximum likelihood (ML) phylogenetic trees, sequences were aligned by MAFFT v7.480 (Katoh et al., 2002), then poorly aligned regions were trimmed by trimAL v1.4 (Capella-Gutiérrez et al., 2009) with the option "-automated1". The alignment files were then subjected to IQ-TREE v1.6.12 (Trifinopoulos et al., 2016) with default settings (1000 bootstrap iterations) and with the best-fit substitution model identified by ModelFinder (Kalyaanamoorthy et al., 2017). For the phylogenetic tree based on the whole chloroplast genomes, the large single-copy (LSC) sequences were used for alignment. For the tree based on four intergenic regions, alignment was done separately for each region and then concatenated into one file. The resulting ML trees were visualised in FigTree v1.4.3 (http://tree.bio.ed.ac.uk/software/figtree/) and rooted using *A. thaliana*, and *Vella spinosa*, respectively. Sequence alignments and machine-readable phylogenetic trees are provided in Supporting Information Dataset S1.

2.13 | Synteny analysis

A pairwise homology search was performed using BLASTp v2.9.0+ (Altschul et al., 1990) followed by predicting synteny of genes across all taxa using MCScanX v0.8 (Wang et al., 2012). A heatmap was generated to visualise the percentage of syntenic genes conserved across all taxa. Finally, we assessed whether the assembly quality has confounding effects on synteny between a pair of taxa by computing Pearson correlations (Figures S7 and S8).

3 | RESULTS

3.1 | De novo genome assemblies

This study reported 19 de novo genome assemblies for 19 taxa of 18 Brassiceae species (Table 1). Seven of the 19 assemblies originated from combining long- and linked-read data, while the other 12 were pure linked-read-based assemblies (Table S1). The assembly size ranged from 271.11 Mbp for *B. tourneforti* to 884.18 Mbp for *D. acris*. The number of scaffolds ranged from 839 for *D. muralis* to 62 651 for *D. harra*. The assembly quality measurements, such as L50 values, ranged from 7247 scaffolds for *B. gravinae* to 7 for *B. tournefortii*, while N50 values ranged from 16.5 Kb for *B. gravinae* to 14.4 Mbp for *B. tourneforti* (Table 1). The genome completeness, assessed by BUSCO against the eudicot_db10 database, ranged from 91% for *B. gravinae* to 99% for *D. muralis*. Except for the allotetraploid *D. muralis*, all assemblies had duplication levels between 12% and 28.1% (Figure 1 and Table S6). In addition, for downstream analyses, the 19 assemblies generated in this study for taxa from the Brassiceae tribe were complemented with 13 publicly available genome assemblies originating from the Brassicaceae family (Table S2). Similarly to our de novo assemblies, the fragmentation level and gene completeness varied across the 13 literature assemblies (Table S7). The recently released assemblies of *M. arvensis* and *M. moricandioides* were polished with the novel linked reads of the same accession, sequenced during this study, resulting, respectively, in 326521 and 231821 single- and multi-nucleotide changes. The purging and scaffolding of the same assemblies using the linked reads improved contiguity marginally without sacrificing BUSCO gene completeness (Table S11).

TABLE 1 Summary statistics of genome assemblies developed in our study along with CO_2 compensation point (CPP) and inferred photosynthesis type from Schlüter et al. (2023).

Species	CPP	Photosynthesis	Scaffolds	Assembly size (bp)	N50 (bp)	L50 (bp)	BUSCO complete	Gene models
B. gravinae	37.22	C3-C4	49 186	448 675 291	16 594	7247	91%	33 701
B. repanda	55.60	C3	15 286	616 556 673	594 246	231	97%	44 116
B. tournefortii	47.46	C3	933	271 107 445	14 406 319	7	98%	34 193
C. annua	55.33	C3	3696	483 163 416	4 251 916	36	98%	37 608
D. acris	55.57	C3	56 348	884 178 440	93 034	2082	97%	84 702
D. erucoides	30.30	C3-C4	4600	359 627 687	1 993 050	46	98%	44 701
D. harra	53.25	C3	62 651	536 054 563	77 558	1388	91%	55 864
D. muralis	35.04	C3-C4	839	435 802 975	3 038 157	38	99%	84 301
D. tenuifolia	12.41	C3-C4	6128	552 030 198	1 665 732	85	97%	44 007
D. tenuisiliqua	49.38	C3	16 194	510 911 626	328 653	315	97%	40 352
D. viminea	51.08	C3	956	304 993 883	3 734 896	22	98%	38 868
E. sativa	51.82	C3	1371	595 848 844	1 142 695	120	97%	46 211
H. incana HIR1	50.50	C3	6478	371 040 007	885 325	73	97%	43 237
H. incana HIR3	38.37	C3-C4	3768	347 540 957	662 613	120	97%	42 185
M. nitens	21.04	C3-C4	28 171	516 121 665	48 092	3575	93%	41 068
M. sinaica	23.90	C3-C4	44 586	825 206 170	195 836	1041	93%	27 349
M. spinosa	17.80	C3-C4	38 590	443 735 849	35 353	1831	94%	63 573
M. suffruticosa	24.87	C3-C4	9798	516 708 898	991 320	120	97%	41 892
S. alba	49.96	C3	13 204	323 376 403	605 444	77	97%	36 457

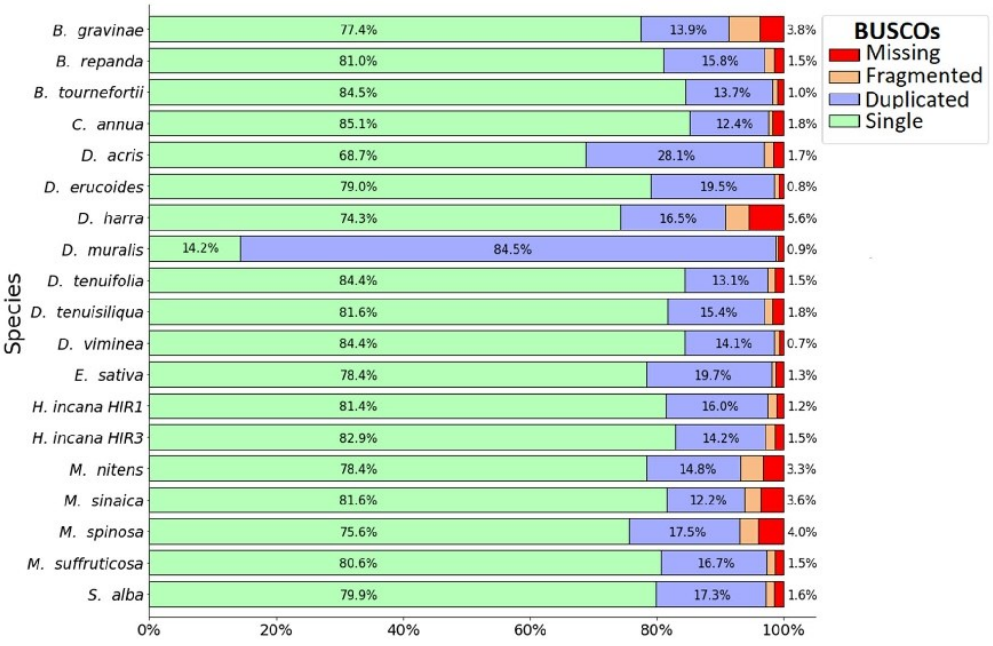

FIGURE 1 Assessment of the completeness of the 19 de novo genome assemblies using BUSCO with eudicot_10db database.

3.2 | Genome size estimation

We estimated the genome size for all taxa included in this study using a k-mer based approach. The genome size estimation failed for *D. acris* (Table S5). The highest genome size estimation was observed for *D. muralis* and the highest heterozygosity level for *M. spinosa*. Our genome size estimates were 50.8% to 85.3% smaller than reported in the literature, with the largest deviations observed for *B. tourneforti* and *H. incana* HIR3.

3.3 | Ploidy estimation

The ploidy of the 19 new assemblies was estimated with nQuire based on the frequency distribution of biallelic variant sites of reads that mapped to BUSCO genes. The resulting estimations were diploid for all assemblies except for *M. spinosa*, for which the estimation was tetraploid, and for *D. harra*, *B. tournefortii*, *D. muralis* and *D. viminea*, where the results were unclear (Figure S1).

3.4 | Repeat annotation

Repeat content analysis identified an average of 1937 unique interspersed repeat families in all genome assemblies, including the ones publicly available. The number of notable repeat families ranged from 171 for *Raphanus raphanistrum* to 2810 for *M. arvensis*. Further, an average of 43% of the respective genome assemblies were masked for gene annotation. The lowest proportion of genome assembly was masked for *R. raphanistrum* (6.35%), while the highest amount of genome assembly was masked for *G. gynandra* (65.22%) (Table S8).

3.5 | Gene annotation

The de novo gene annotation was performed for all 19 genome assemblies from this study as well as the publicly available genome assemblies of 13 species using the same method to facilitate comparisons across taxa. The annotations produced a median of 45 408 gene models per assembly, ranging from 22 318 gene models for *A. thaliana* to 113 686 gene models for *B. napus* (Table S9) with an average annotation edit distance (AED) score of 0.186. The annotations with the lowest cumulative AED scores were that of *D. acris* and *M. spinosa* (Figure 2). In contrast, the annotations with the highest cumulative AED score were that of *M. moricandioides*, *M. sinaica* and *Sinapis alba* (Figure 2). An average of 3648 gene models per assembly were discarded due to high AED scores (>0.5) or their small size (<50 amino acids). In addition, an average of 1937 gene models per assembly were discarded due to their functional annotations related to TEs. The final gene models for each assembly retained an average of 90.9% BUSCO genes, except for *B. gravinae* (65%) (Figure S2). In addition, gene length distribution analysis revealed that the mean and median gene lengths were 1880 bp and 1441 bp across taxa, respectively (Figure S3). In contrast, the bigger difference between the mean and median inter-genic distances

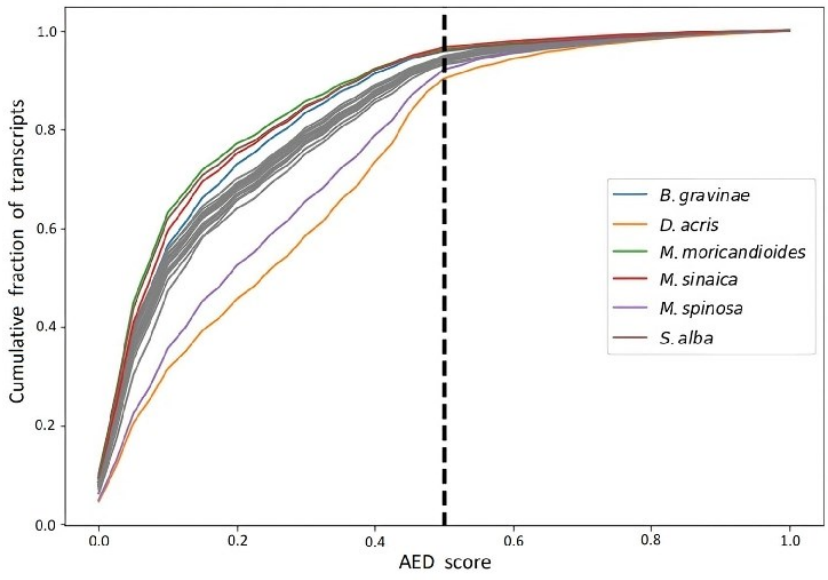

FIGURE 2 Cumulative annotation edit distance (AED) score of gene annotations for the 19 de novo assemblies and 13 publicly available genome assemblies. Highlighted in colour are the assemblies with the highest and lowest proportion of quality gene annotations (AED < 0.5). [Color figure can be viewed at wileyonlinelibrary.com]

(6662 bp vs. 2176 bp) compared to the gene length suggested a more skewed distribution of the former (Figure S4).

The length of the upstream sequences for all genes were measured in each assembly by measuring sequence length from the transcription start site (TSS) until interruption due to contig end or insertion of Ns. We found that the assemblies of *B. gravinae*, *D. harra* and *M. sinaica* contained a high percentage of genes with short upstream sequences (<1 Kb). In contrast, all other assemblies were characterised by availability of very long (>30 Kb) upstream sequences for the majority of genes (Figure S5).

3.6 | Orthology

Orthofinder clustered about 98% of all protein sequences (1 372 043) from all 32 assemblies into 42 928 orthogroups (HOGs) (Table S10). Each HOG indicates a set of homologous genes descended from all taxa's last common ancestor gene (Emms & Kelly, 2019). Of the 42 928 HOGs, 22 694 were present in less than 10 assemblies and were filtered out to avoid potential biases when creating the phylogenetic tree. After this filtering step, most assemblies had a median of one gene per HOG (Figure 3a). Exceptions with a higher median of two or three were observed for the tetraploid species and some diploid species (Figure 3a), all of which contained a higher number of total genes (Figure 3b). The percentage of single-copy genes varied from 5.9% in *A. thaliana* to 60.1% in *B. napus* (Figure 3b). Most commonly, HOGs existed in (a) all assemblies; (b) all but one assembly; (c) *D. muralis* and *D. tenuifolia* or *D. viminea*; (d) *B. napus* and *B. oleracea* (Figure S6).

3.7 | Genome-wide phylogenetic tree

We established a genome-wide phylogenetic tree for all species included in this study by using ASTRAL-pro with 27 793 HOGs. The phylogenetic tree contained two main clades (Figure 4): one clade comprising the taxa of the Moricandia genus and most taxa of the Diplotaxis genus, as well as *E. sativa*. In contrast, the other clade contained several Brassica taxa and the Raphanus and Hirschfeldia genera. As outgroup, we used *G. gynandra* from the Cleomaceae family, which diverged from the Brassicaceae approximately 40 million years ago (Edger et al., 2018). The Moricandia genus was monophyletic, while the Diplotaxis and Brassica genera species were dispersed across the phylogenetic tree (Figure 4). When integrating the information of the CO_2 compensation points from Schlüter et al. (2023) in the phylogenetic tree, the result indicated that the C3-C4 intermediate photosynthesis might have developed five times independently (Figure 4).

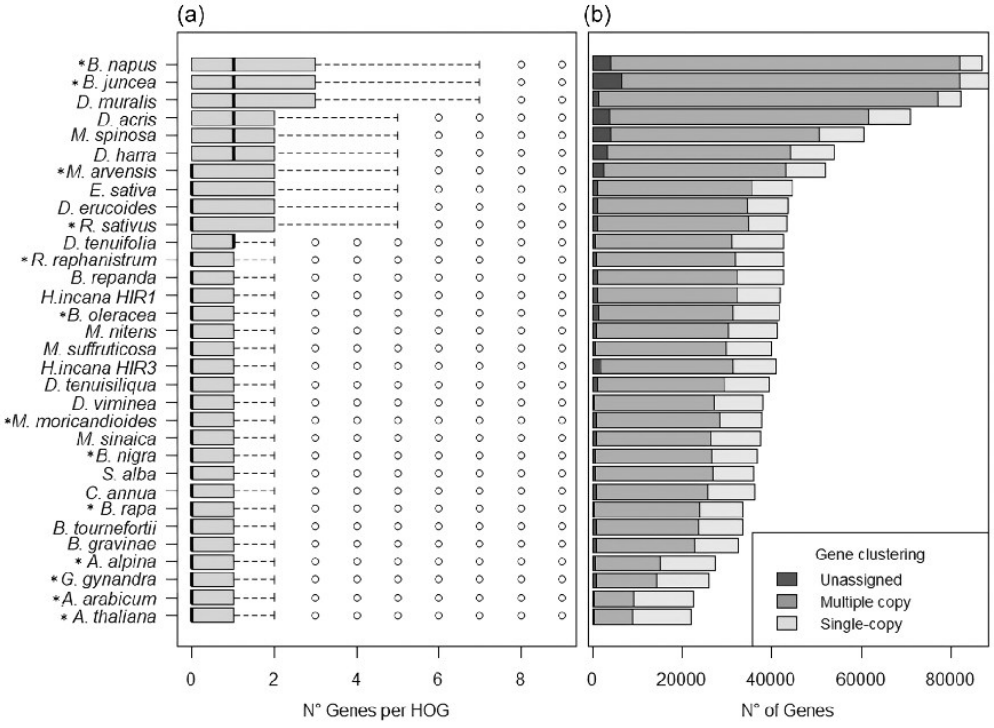

FIGURE 3 (a) Distribution of gene numbers per hierarchical orthogroup (HOG) per taxa in all 32 assemblies (b) the total number of unassigned genes, assigned together with other copies or as a single copy to an orthogroup. Species with previously available genome assemblies are marked with an asterisk (*).

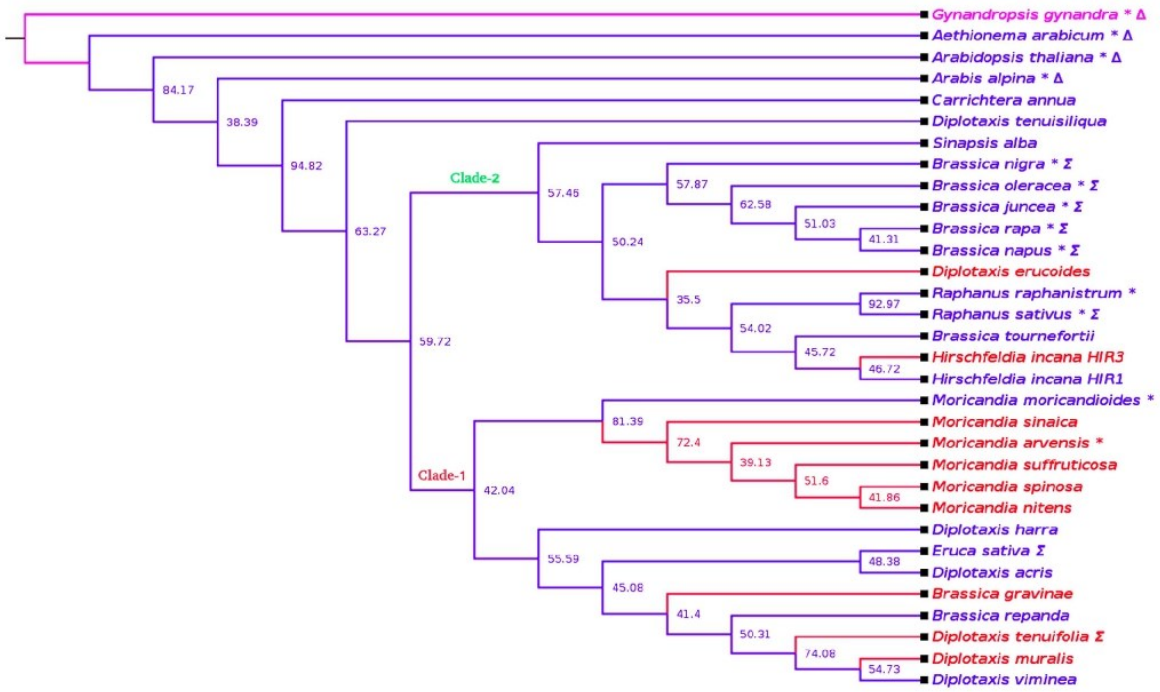

FIGURE 4 Species tree created using a multi-species coalescent-based approach with *G. gynandra* (C4 photosynthesis) as an outgroup species. Node values are quartet scores created by Astral-Pro, indicating branching support on a 0–100 range, representing the percentage of quartets in the gene trees that agree with the branch in the species tree. The colour code indicates the photosynthesis type inferred from CPP values (Schlüter et al., 2023): Blue colour indicates C3 photosynthesis, while red colour indicates C3-C4 photosynthesis. Species with previously available genome assemblies are marked with an asterisk (*). Model and crop species are marked with a triangle (Δ) or sigma (∑), respectively. [Color figure can be viewed at wileyonlinelibrary.com]

3.8 | Synteny map

We created a synteny map for all 32 assemblies by computing pairwise collinear genes and observed a high conservation of syntenic genes between most taxa. Of particular interest was the high extent of synteny of *D. muralis* with *D. tenuifolia* as well as *D. viminea*. In contrast, *B. gravinae*, *D. acris*, *D. harra*, *M. sinaica* and *R. raphanistrum* showed a very low synteny against all other taxa (Figure 5). Therefore, we performed a correlation analysis to quantify the influence of assembly quality measured as the number of scaffolds (i.e., assembly fragmentation) on the conservation of synteny between each pair of taxa (Figure S7). A negative correlation (−0.503 or −0.739 for the number or percentage of collinear genes) between assembly quality (fragmentation) and synteny was observed, indicating that the above reported low synteny can be explained by differences in assembly quality. However, a linear regression considering both phylogenetic branch lengths and the sum of N90 scaffolds between each pair of species also marked phylogenetic branch length as significant factor in relation to synteny. That relation is negative, meaning that distantly related species tend to share less synteny than closely related ones (Table S12).

3.9 | Phylogenetic analysis of *H. incana* accessions based on chloroplast sequences

To resolve the phylogenetic relationships of the two *H. incana* accessions HIR1 and HIR3 from our study and another accession, NIJ, previously reported in Garassino et al. (2022), we constructed phylogenetic trees. Our phylogenetic tree based on the LSC regions of 14 Brassiceae chloroplast genomes had high support values for all nodes (SH-aLRT/bootstrap >70%). The three *H. incana* accessions were placed together on the same branch, next to a sister branch comprising *B. nigra* and *S. alba* within the Nigra clade (SH-aLRT/bootstrap = 100%) (Figure 6a). To further resolve the relationship amongst these three accessions, we constructed another phylogenetic tree based on four chloroplast intergenic regions in which we included more closely related species *Erucastrum virgatum*, *B. procubens*, *S. pubescens*, *B. tournefortii* and another *H. incana* accession, BGV UPM, from Arias & Pires, (2012). These analyses suggest that *H. incana* HIR3 is genetically distant from HIR1, NIJ, and BGV UPM accessions (Figure 6b) but is located close to *S. pubescens* and *B. procubens* on one branch, whereas the other three accessions are located on another branch with *E. virgatum* (SH-aLRT/bootstrap >70%). The differences were further supported by plant morphology

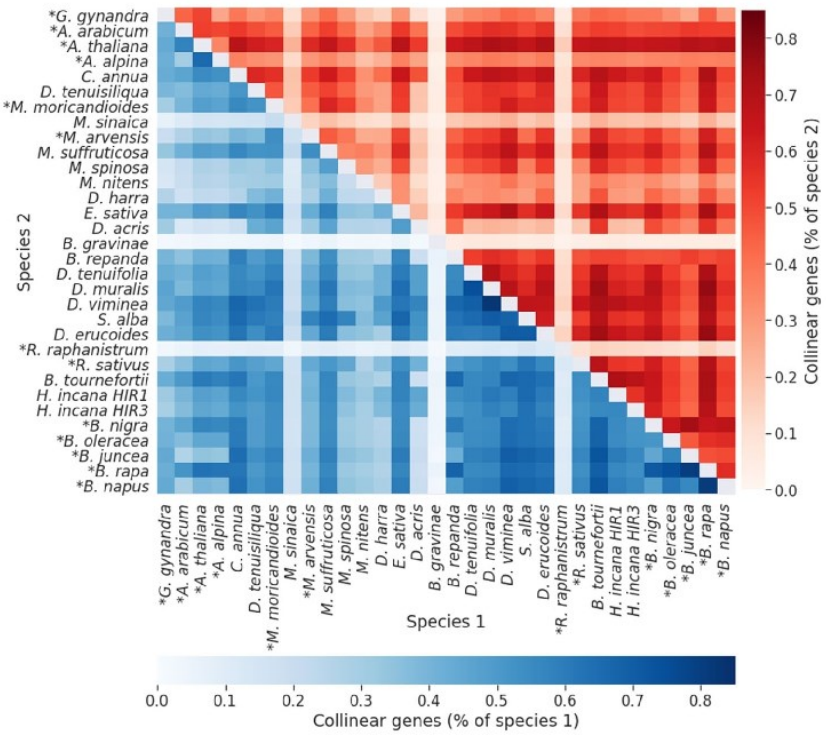

FIGURE 5 Heatmap of the percentage of syntenic genes between each pair of species with species 1 as reference (below the diagonal) and species 2 as reference (above the diagonal). The species were sorted according to their position in the phylogenetic tree. Species with previously available genome assemblies are marked with an asterisk (*). [Color figure can be viewed at wileyonlinelibrary.com]

of the three *H. incana* accessions NIJ, HIR1 and HIR3 (Figures 6c and S9).

4 | DISCUSSION

The main aim of this study was to establish resources that enable genomic comparisons and investigations of the evolution of C3-C4 intermediate photosynthesis within the Brassicaceae family and especially the Brassiceae tribe. For such analyses, not only dense sampling of C3-C4 intermediate species but also of closely related C3 species is required, to be able to separate signal from noise. Therefore, we have included in this study all species of the Brassiceae tribe whose genome was not yet sequenced and for which we were able to obtain seeds.

4.1 | High-quality draft de novo genome assemblies and annotations

Overall, the genome assemblies generated in this study were fragmented, with varying levels of contiguity and quality. Nevertheless, with at least 91% complete genes identified by BUSCO, our assemblies captured most of the gene space (Figure 1), which indicates suitability for comparative genomic analysis. The duplication rates are relatively low, except for *D. muralis* at 84% (discussed below) and *D. acris* at 28% (Figure 1). No k-mer estimation of heterozygosity was possible for *D. acris* (Table S5). Furthermore, all presented assemblies meet the minimum requirement of an N50 larger than average gene length (Yandell & Ence, 2012). Even the N90 values of all assemblies were higher than the average gene length of around 2000 bp (Figures S3 and S4). Therefore, the quality of the genome assemblies in this study is comparable or higher than the assemblies available for other Brassicaceae species with similar genome sizes (e.g., Haudry et al., 2013; Lin et al., 2021; Moghe et al., 2014).

The most contiguous assembly was realised in our study for *D. tenuisiliqua* (Tables S1 and S4). This is presumably due to the two generations of selfing performed before sequencing (cf. Li & Harkess, 2018). In addition, the assembly of *D. muralis* reached satisfactory contiguity after two generations of selfing. Coincidentally, an assembly for *H. incana (Nijmegen)* with six generations of selfing has just been released (Garassino et al., 2022) with an N50 value of 5.1 Mb, which is considerably longer compared to our assemblies with N50s of 885 and 663 Kb for *H. incana* HIR1 and *H. incana* HIR3 accessions, respectively, that were sequenced without prior selfing.

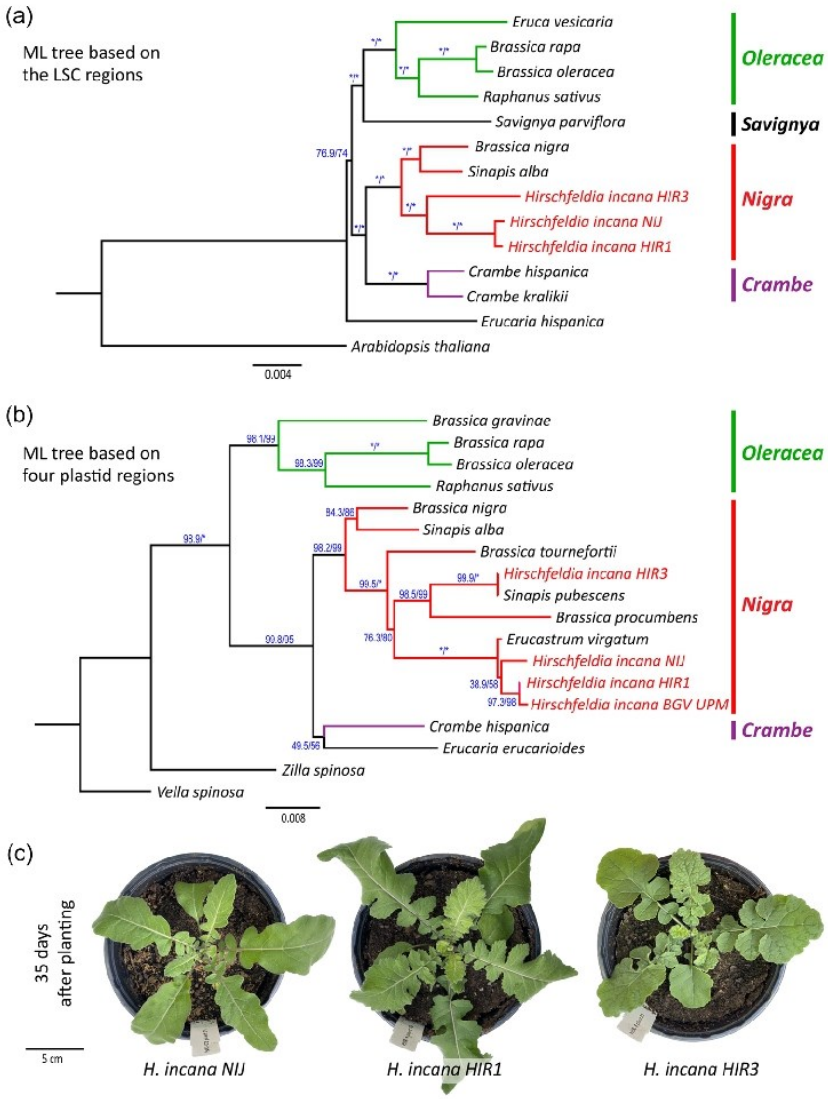

FIGURE 6 Phylogenetic trees of three *H. incana* accessions in relation to other species in Brassiceae. (a) ML phylogeny reconstruction was done using IQ-TREE based on the LSC regions of the chloroplast genomes from 14 species, and (b) four rapidly evolving chloroplast intergenic regions, rpl32-trnL, atpI-atpH, psbD-trnT and ycf6-psbM. Supporting values are SH-aLRT (Shimodaira-Hasegawa-like approximate likelihood ratio) support (%)/ultrafast bootstrap support (%), respectively, and are given next to the branch. An asterisk (*) denotes a supporting value of 100%. Branch length denotes substitutions per site. Trees were rooted (a) using *A. thaliana*, and (b) *Vella spinosa*, respectively. (c) Phenotypes of three *H. incana* accessions, NIJ, HIR1 and HIR3. Photos taken at 35 days after planting. LSC, large single-copy; ML, maximum likelihood. [Color figure can be viewed at wileyonlinelibrary.com]

The assemblies developed in our study result mostly from single linked-read libraries. Contiguity and completeness can be further improved by scaffolding and gap-filling using low-coverage long-read sequencing data, as we illustrated with *H. incana* HIR1, *D. acris*, *D. harra* and *M. sinaica*. In contrast, we improved the long-read assemblies by scaffolding and polishing with linked-read data for *E. sativa* and *D. erucoides*.

Although the assemblies created in this study are scaffold-level, the assessment of the lengths of the upstream sequences of the TSSs of annotated genes showed that for a high proportion of annotated genes upstream sequences are available in most of the assemblies (Figure S5). Hence, the dataset generated in this study enables analyses downstream from transcriptomics studies, such as the deep analysis of cis-regulatory motifs. Such an analysis could prove crucial

in advancing the knowledge about differences underlying gene regulation mechanisms between C3 and C3-C4 intermediate species.

Despite the lack of transcriptional data for all 19 taxa, our de novo gene annotation strategy resulted in an average of 46 546 gene models with high quality (AED ≤ 0.5) across taxa (Figure 2 and Table S9). The number of gene models for each taxa was comparable to the number of gene models in the gene annotation of the publicly available 13 species that complemented our study.

4.2 | Conservation of genes across Brassiceae species

Resolving the orthologous relationships between species is fundamental to comparative genomics. We grouped around 98% of annotated genes into orthogroups (Figure 3 and Table S10), where 65.18% were grouped into 11 072 orthogroups present in at least 31 of the used 32 taxa (Figure S6). Notably, some orthogroups were absent in species due to low-quality assembly, as evident with *B. gravinae* (Figure S6 and Table S6), or large phylogenetic distances, as evident from *G. gynandra* and potentially with *Aethionema arabicum* (Figure 4). In contrast, some orthogroups were present exclusively in a specific taxa (Figure S6 and Table S10). Moreover, these species-specific orthogroups often contained TE-related genes (Jayakodi et al., 2020).

To illustrate the possibility of comparing gene regions between pairs of species, we generated a synteny map. As previously mentioned, we consider the particularly high synteny values for *D. muralis*, *D. tenuifolia* and *D. viminea* (Figure 5) as evidence of its hybrid origin. Additional high conservation of synteny is visible across many taxa, with exceptions for *B. gravinae*, *D. acris*, *D. harra*, *M. sinaica* and *R. raphanistrum* (Figure 5). These species had lower assembly contiguity, which correlated strongly with synteny of genes between species (Figure S7). When disconsidering these species, the Pearson correlation coefficient between synteny colinear gene number decreased from −0.503 values to −0.069 (Figure S8), which indicates that the quality of all other annotations is at a more comparable standard and, thus, can be used for comparative genomics projects that focus on the Brassicaceae family.

4.3 | Interspecific hybridisation in Diplotaxis and Moricandida

It is thought that *D. muralis* and *M. spinosa* are derived from past hybridisation events between *D. tenuifolia* and *D. viminea* (Ueno et al., 2006) and between *M. suffruticosa* and *M. nitens* (Perfectti et al., 2017), respectively. The placement of both species close to the respective parental species in the phylogenetic tree (Figure 5) and the support for tetraploidy by nQuire in *M. spinosa* (Figure S1) support these ideas. Further support for tetraploidy in *D. muralis* was found in its large estimated genome size (Table S5), high synteny with both parental species (Figure 5) and by the large number of HOGs containing only *D. muralis* and either of the parent species (Figure S6). Therewith, our work constitutes the first genomic support for the hybrid hypothesis in *D. muralis*, whereas previous support came from isoenzyme pattern and random amplified polymorphic DNA (Eschmann-Grupe et al., 2004). However, the same conclusions cannot be clearly applied to *M. spinosa*. Namely, the smaller genome size estimation and the very high heterozygosity (Table S5) lead us to speculate that it could be an autopolyploid closely related but not derived from *M. nitens* and *M. suffruticosa* hybridisation. It is worth considering that sequence divergence between *M. suffruticosa* and *M. nitens* might be much smaller than between *D. tenuifolia* and *D. viminea*, in which case a hybrid would look more like an auto-polyploid. However, *M. spinosa* shared only a moderate amount of HOGs in exclusivity with *M. nitens*, and a small amount with *M. suffruticosa* (Figure S6). Therefore, we recommend the estimation of sequence identity between the genomes and divergence time with molecular clocks.

4.4 | Evolution of C3-C4 intermediate photosynthesis in the Brassiceae species

Resolving phylogenetic relationships between species is fundamental to evolutionary analysis, which provides a framework to explore the evolution of traits across species. Therefore, we estimated a phylogenetic tree using a multi-species coalescent-based approach (Flouri et al., 2018) to understand how the C3-C4 intermediate photosynthetic trait evolved in the Brassiceae tribe (Figure 5). The relative placement of *H. incana*, *R. sativus*, and *S. alba* observed in our study agrees with the literature (Huang et al., 2016), whereas the placement of *B. tournefortii* within this clade has not been described earlier. More interestingly, the phylogenetic tree indicates that the C3-C4 intermediate photosynthesis may have evolved independently up to five times in the Brassiceae tribe. Since the Brassiceae does not contain bona fide C4 species (Sage et al., 2011), we consider it unlikely that the C3-C4 intermediate trait in this tribe has evolved through the hybridisation of a C3 and a C4 species (Kadereit et al., 2017). However, we cannot exclude that Brassiceae at some point in time contained C4 species that went extinct.

4.5 | Phylogenetic analysis of *H. incana* accessions based on genome and chloroplast sequences

The HIR3 and HIR1 accessions of *H. incana* were placed as sister species in the phylogenetic tree derived from genome-wide sequences (Figure 5). As these differ considerably in their photosynthetic properties (i.e., C3 vs. C3-C4) additional phylogenetic analyses were performed. Our phylogenetic tree based on the LSC regions of 14 Brassiceae chloroplast genomes was consistent with the topology reported in previous studies (Arias & Pires, 2012; Koch & Lemmel, 2019). In this tree, the two *H. incana* accessions were placed together with the NIJ accession (Garassino et al., 2022) on the same branch. However, our result suggests that HIR3 is genetically different from the HIR1 and NIJ

accessions (Figure 6a). The relationships of these accessions were further resolved by a phylogenetic tree based on four chloroplast intergenic regions of species that clustered closely to *H. incana* in Arias & Pires, (2012). This tree revealed that HIR3 is located close to *S. pubescens* and *B. procubens* on one branch, whereas all previously identified *H. incana* accessions including HIR1 and NIJ are located on another branch with *E. virgatum* (Figures 6b and S9). Together with the observed morphological differences among the three *H. incana* accessions (Figures 6c and S9), our results suggested that HIR3 belongs to a different species than *H. incana*.

4.6 | Further directions in researching C3-C4 photosynthesis in the Brassiceae tribe

Most earlier literature comparing C3-C4 photosynthesis within the Brassiceae has focused on the Moricandia and Diplotaxis genera (Adwy et al., 2015; Razmjoo et al., 1996; Schlüter et al., 2017; Ueno et al., 2006). Both belong to a separate subclade where the species with highest commercial value is rocket salad *E. sativa* (Figure 5). However, the phylogenetic tree of our study indicates the existence of two C3-C4 intermediate species and taxa, namely *D. erucoides* and HIR3, in the same subclade where the commercially important species of the Brassica and Raphanus genera reside (Figure 4). Therefore, *D. erucoides* and HIR3 might be appropriate sources to transfer photosynthetic properties of C3-C4 species to the Brassica and Raphanus crops by establishing interspecific crosses. Furthermore, such approaches will be facilitated by now possible detailed comparative genomic studies between the C3-C4 species *D. erucoides* and HIR3 with the Brassica and Raphanus crops but as well as with the currently known closest C3 relatives *H. incana* HIR1 or *B. tournefortii*. Finally, while pairwise comparisons between close C3 and C3-C4 relatives can yield first insights into the genomic and, thus, physiological differences between those species, a more holistic approach comparing multiple taxa and considering their evolutionary distance is required for a genetic dissection of the interspecific differences with a high statistical power (Nagy et al., 2020). Such analyses could be performed with our panel of species using the phylogenetic association mapping framework described and used earlier (Hiller et al., 2012; Kiefer et al., 2019; Prudent et al., 2016; Smith et al., 2020). This framework tests e.g. in a mixed-model for the significance of associations between any genomic variant and phenotypic differences, such as in our context CO_2 compensation point while controlling for phylogenetic distances. This has the potential to identify common genetic factors in our species panel that are responsible for differences in the CO_2 compensation point that are not just coincident to one lineage.

5 | CONCLUSION

We generated draft de novo genome assemblies using linked- and long-read sequencing data for 19 taxa of the Brassiceae tribe, doubling the sampling depth of genomes within this tribe. Our gene annotation generated high quality models as well as potential to explore variants in genes and regulatory sequences, while our phylogenetic tree indicates that intermediate C3-C4 photosynthesis evolved five times independently across these taxa. This work constitutes the first genomic evidence that *D. muralis* is a hybrid of *D. tenuifolia* and *D. viminea*, and that the HIR3 accession of *H. incana* is a separate species from other studied accessions, having C3-C4 characteristics. Altogether, the high-quality de novo genome assemblies and the gene annotation will be helpful to the scientific community in exploring further the evolution of C3-C4 intermediate photosynthesis in the Brassiceae tribe.

ACKNOWLEDGEMENTS

The authors give thanks to the Millenium Seed Bank Kew Gardens and the federal ex situ gene bank Gatersleben for proving seeds of the species and taxa used in this study. Computational infrastructure and support were provided by the Centre for Information and Media Technology at Heinrich Heine University Düsseldorf. Sequencing support was provided by the Genomics & Transcriptomics Laboratory (GTL) of the Heinrich Heine University Düsseldorf as part of the West German Genome Center (WGGC) and by Max Planck-Genome-Centre Cologne (MP-GC). We thank our colleagues Stephanie Krey and Anja Kyriacidis for their excellent technical assistance. This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) in the frame of the ERA-CAPS project C4BREED, Germany's Excellence Strategy (EXC 2048/1, Project ID: 390686111), and a Collaborative Research Centre/Transregio (TRR 341, Project ID: 456082119). Open Access funding enabled and organized by Projekt DEAL.

DATA AVAILABILITY STATEMENT

The raw Sequence Read Archive data used in this study is deposited in NCBI under BioProject PRJNA905373. The scripts used for genome assembly, annotation and orthology identification were uploaded to github (https://github.com/ViriatoII/C4Evol). For convenience when using orthogroups of our annotated genes, we publish a Python tool that annotates orthogroups with the gene names of a reference annotation (E.g. *A. thaliana*), as well as GO annotations and PFAM domains, using blast and interproscan (https://github.com/ViriatoII/C4Evol/blob/master/annotate_ogroups_vs_ref.py). The final assemblies and annotation as well as the Supplementary dataset 1 have been uploaded to Figshare: (https://figshare.com/articles/dataset/C4Evol_Brassicaceae_genomes/21671201).

ORCID

Ricardo Guerreiro http://orcid.org/0000-0003-2472-5260
Venkata Suresh Bonthala http://orcid.org/0000-0001-6550-1648
Nam V. Hoang http://orcid.org/0000-0003-0782-2835
Benjamin Stich http://orcid.org/0000-0001-6791-8068

REFERENCES

Adwy, W., Laxa, M. & Peterhansel, C. (2015) A simple mechanism for the establishment of C2-specific gene expression in Brassicaceae. *The*

Plant Journal, 84, 1231–1238. Available at https://doi.org/10.1111/tpj.13084

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, 100176. Available at https://doi.org/10.1016/j.plgene.2019.100176

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) Basic local alignment search tool. *Journal of Molecular Biology*, 215(3), 403–410.

Apel, P., Horstmann, C. & Pfeffer, M. (1997) The Moricandia syndrome in species of the Brassicaceae-evolutionary aspects. *Photosynthetica*, 33(2), 205–215.

Arias, T. & Pires, J.C. (2012) A fully resolved chloroplast phylogeny of the brassica crops and wild relatives (Brassicaceae: Brassiceae): novel clades and potential taxonomic implications. *Taxon*, 61(5), 980–988.

Arseneau, J.-R., Steeves, R. & Laflamme, M. (2017) Modified low-salt CTAB extraction of high-quality DNA from contaminant-rich tissues. *Molecular Ecology Resources*, 17(4), 686–693. Available at https://doi.org/10.1111/1755-0998.12616

Barrell, D., Dimmer, E., Huntley, R.P., Binns, D., O'Donovan, C. & Apweiler, R. (2009) The GOA database in 2009—an integrated gene ontology annotation resource. *Nucleic Acids Research*, 37(Database), D396–D403. Available at https://doi.org/10.1093/nar/gkn803

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*, 223(1), 150–166. Available at https://doi.org/10.1111/NPH.15787

Bolger, A.M., Lohse, M. & Usadel, B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15), 2114–2120. Available at https://doi.org/10.1093/BIOINFORMATICS/BTU170

Bräutigam, A., Kajala, K., Wullenweber, J., Sommer, M., Gagneul, D., Weber, K.L. et al. (2011) An mRNA blueprint for C4 photosynthesis derived from comparative transcriptomics of closely related C3 and C4 species. *Plant Physiology*, 155(1), 142–156. Available at https://doi.org/10.1104/pp.110.159442

von Caemmerer, S., Quick, W.P. & Furbank, R.T. (2012) The development of C4 rice: current progress and future challenges. *Science*, 336(6089), 1671–1672.

Campbell, M.S., Holt, C., Moore, B. & Yandell, M. (2014) Genome annotation and curation using MAKER and MAKER-P. In, *Current Protocols in Bioinformatics*, Vol. 48(Issue 1). John Wiley & Sons, Inc. pp. 4.11.1–4.11.39. https://doi.org/10.1002/0471250953.bi0411s48

Capella-Gutiérrez, S., Silla-Martínez, J.M. & Gabaldón, T. (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, 25(15), 1972–1973.

Cheng, C.-Y., Krishnakumar, V., Chan, A.P., Thibaud-Nissen, F., Schobel, S. & Town, C.D. (2017) Araport11: a complete reannotation of the *Arabidopsis thaliana* reference genome. *The Plant Journal*, 89(4), 789–804. Available at https://doi.org/10.1111/tpj.13415

Conant, G.C., Birchler, J.A. & Pires, J.C. (2014) Dosage, duplication, and diploidization: clarifying the interplay of multiple models for duplicate gene evolution over time. *Current Opinion in Plant Biology*, 19, 91–98. Available at https://doi.org/10.1016/J.PBI.2014.05.008

De Coster, W., D'hert, S., Schultz, D.T., Cruts, M. & Van Broeckhoven, C. (2018) NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics*, 34(15), 2666–2669.

Denton, A.K., Maß, J., Külahoglu, C., Lercher, M.J., Bräutigam, A. & Weber, A.P.M. (2017) Freeze-quenched maize mesophyll and bundle sheath separation uncovers bias in previous tissue-specific RNA-Seq data. *Journal of Experimental Botany*, 68(2), 147–160.

Edger, P.P., Hall, J.C., Harkess, A., Tang, M., Coombs, J., Mohammadin, S. et al. (2018) Brassicales phylogeny inferred from 72 plastid genes: a reanalysis of the phyflogenetic localization of two paleopolyploid events and origin of novel chemical defense. *American Journal of Botany*, 105(3), 463–469. Available at https://doi.org/10.1002/ajb2.1040s

Edwards, G.E. & Ku, M.S.B. (1987) Biochemistry of C3–C4 intermediates, *Photosynthesis*. Elsevier. pp. 275–325.

Eisenhut, M., Roell, M.S. & Weber, A.P.M. (2019) Mechanistic understanding of photorespiration paves the way to a new green revolution. *New Phytologist*, 223(4), 1762–1769.

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S.R., Luciani, A., Potter, S.C. et al. (2019) The Pfam protein families database in 2019. *Nucleic Acids Research*, 47(D1), D427–D432. Available at https://doi.org/10.1093/nar/gky995

Ellinghaus, D., Kurtz, S. & Willhoeft, U. (2008) LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics*, 9(1), 18.

Emms, D.M. & Kelly, S. (2015) OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology*, 16(1), 157.

Emms, D.M. & Kelly, S. (2019) OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology*, 20(1), 238. Available at https://doi.org/10.1186/s13059-019-1832-y

Ermakova, M., Arrivault, S., Giuliani, R., Danila, F., Alonso-Cantabrana, H., Vlad, D. et al. (2021) Installation of C4 photosynthetic pathway enzymes in rice using a single construct. *Plant Biotechnology Journal*, 19(3), 575–588. Available at https://doi.org/10.1111/PBI.13487

Eschmann-Grupe, G., Neuffer, B. & Hurka, H. (2004) Extent and structure of genetic variation in two colonising Diplotaxis species (Brassicaceae) with contrasting breeding systems. *Plant Systematics and Evolution*, 244(1), 31–43.

Flouri, T., Jiao, X., Rannala, B. & Yang, Z. (2018) Species tree inference with BPP using genomic sequences and the multispecies coalescent. *Molecular Biology and Evolution*, 35(10), 2585–2593. Available at https://doi.org/10.1093/molbev/msy147

Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28, 3150–3152. Available at https://doi.org/10.1093/bioinformatics/bts565

Garassino, F., Wijfjes, R.Y., Boesten, R., Reyes Marquez, F., Becker, F.F.M., Clapero, V. et al. (2022) The genome sequence of *Hirschfeldia incana*, a new Brassicaceae model to improve photosynthetic light-use efficiency. *The Plant Journal*, 112(5), 1298–1315.

Gowik, U., Bräutigam, A., Weber, K.L., Weber, A.P.M. & Westhoff, P. (2011) Evolution of C4 photosynthesis in the genus Flaveria: how many and which genes does it take to make C4? *The Plant Cell*, 23, 2087–2105. Available at https://doi.org/10.1105/tpc.111.086264

Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J. et al. (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, 8(8), 1494–1512. Available at https://doi.org/10.1038/nprot.2013.084

Han, Y. & Wessler, S.R. (2010) MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Research*, 38(22), e199.

Hatch, M.D. (1987) C4 photosynthesis: a unique elend of modified biochemistry, anatomy and ultrastructure. *Biochimica et Biophysica Acta (BBA) - Reviews on Bioenergetics*, 895(2), 81–106. Available at https://doi.org/10.1016/S0304-4173(87)80009-5

Haudry, A., Platts, A.E., Vello, E., Hoen, D.R., Leclercq, M., Williamson, R.J. et al. (2013) An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. *Nature Genetics*, 45(8), 891–898. Available at https://doi.org/10.1038/ng.2684

Hiller, M., Schaar, B.T., Indjeian, V.B., Kingsley, D.M., Hagey, L.R. & Bejerano, G. (2012) A "forward genomics" approach links genotype to phenotype using independent phenotypic losses among related species. *Cell Reports*, 2(4), 817–823.

Hoang, N.V., Sogbohossou, E.O.D., Xiong, W., Simpson, C.J.C., Singh, P., Walden, N. et al. (2023) The *Gynandropsis gynandra* genome provides insights into whole-genome duplications and the evolution of C4 photosynthesis in Cleomaceae. *The Plant Cell*, 35, 1334–1359. Available at https://doi.org/10.1093/plcell/koad018

Hoff, K.J. & Stanke, M. (2019) Predicting genes in single genomes with AUGUSTUS. *Current Protocols in Bioinformatics*, 65(1), 57. Available at https://doi.org/10.1002/CPBI.57

Huang, C.-H., Sun, R., Hu, Y., Zeng, L., Zhang, N., Cai, L. et al. (2016) Resolution of Brassicaceae phylogeny using nuclear genes uncovers nested radiations and supports convergent morphological evolution. *Molecular Biology and Evolution*, 33(2), 394–412.

Jackman, S.D., Coombe, L., Chu, J., Warren, R.L., Vandervalk, B.P., Yeo, S. et al. (2018) Tigmint: correcting assembly errors using linked reads from large molecules. *BMC Bioinformatics*, 19(1), 393. Available at https://doi.org/10.1186/s12859-018-2425-6

Jayakodi, M., Padmarasu, S., Haberer, G., Bonthala, V.S., Gundlach, H., Monat, C. et al. (2020) The barley pan-genome reveals the hidden legacy of mutation breeding. *Nature*, 588(7837), 284–289.

Jeong, S.Y., Ahn, H., Ryu, J., Oh, Y., Sivanandhan, G., Won, K.-H. et al. (2019) Generation of early-flowering Chinese cabbage (*Brassica rapa* spp. pekinensis) through CRISPR/Cas9-mediated genome editing. *Plant Biotechnology Reports*, 13(5), 491–499.

Jin, J.-J., Yu, W.-B., Yang, J.-B., Song, Y., dePamphilis, C.W., Yi, T.-S. et al. (2020) GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. *Genome Biology*, 21(1), 241.

Kadereit, G., Bohley, K., Lauterbach, M., Tefarikis, D.T. & Kadereit, J.W. (2017) C3–C4 intermediates may be of hybrid origin – a reminder. *New Phytologist*, 215(1), 70–76. Available at https://doi.org/10.1111/NPH.14567

Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., von Haeseler, A. & Jermiin, L.S. (2017) ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods*, 14(6), 587–589.

Kaneko, Y. & Bang, S.W. (2014) Interspecific and intergeneric hybridization and chromosomal engineering of Brassicaceae crops. *Breeding Science*, 64(1), 14–22.

Katoh, K. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, 30(14), 3059–3066.

Khoshravesh, R., Stinson, C.R., Stata, M., Busch, F.A., Sage, R.F., Ludwig, M. et al. (2016) C3–C4 intermediacy in grasses: organelle enrichment and distribution, glycine decarboxylase expression, and the rise of C2 photosynthesis. *Journal of Experimental Botany*, 67(10), 3065–3078. Available at https://doi.org/10.1093/JXB/ERW150

Kiefer, C., Willing, E.M., Jiao, W.B., Sun, H., Piednoël, M., Hümann, U. et al. (2019) Interspecies association mapping links reduced CG to TG substitution rates to the loss of gene-body methylation. *Nature Plants*, 5(8), 846–855. Available at https://doi.org/10.1038/s41477-019-0486-9

Koch, M.A. & Lemmel, C. (2019) Zahora, a new monotypic genus from tribe Brassiceae (Brassicaceae) endemic to the Moroccan Sahara. *PhytoKeys*, 135, 119–131.

Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H. & Phillippy, A.M. (2017) Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research*, 27(5), 722–736. Available at https://doi.org/10.1101/gr.215087.116

Korf, I. (2004) Gene finding in novel genomes. *BMC Bioinformatics*, 5(1), 59. Available at https://doi.org/10.1186/1471-2105-5-59

Ku, M.S.B., Wu, J., Dai, Z., Scott, R.A., Chu, C. & Edwards, G.E. (1991) Photosynthetic and photorespiratory characteristics of Flaveria species. *Plant Physiology*, 96(2), 518–528.

Külahoglu, C., Denton, A.K., Sommer, M., Maß, J., Schliesky, S., Wrobel, T.J. et al. (2014) Comparative transcriptome atlases reveal altered gene expression modules between two Cleomaceae C3 and C4 plant species. *The Plant Cell*, 26(8), 3243–3260.

Lauterbach, M., Schmidt, H., Billakurthi, K., Hankeln, T., Westhoff, P., Gowik, U. et al. (2017) De novo transcriptome assembly and comparison of C3, C3-C4, and C4 species of tribe salsoleae (Chenopodiaceae). *Frontiers in Plant Science*, 8. Available at https://doi.org/10.3389/fpls.2017.01939

Lewis, S.E. (2004) Gene Ontology: looking backwards and forwards. *Genome Biology*, 6(1), 103. Available at https://doi.org/10.1186/gb-2004-6-1-103

Li, F.W. & Harkess, A. (2018) A guide to sequence your favorite plant genomes, *In* Applications in Plant Sciences, Vol. 6(Issue 3). John Wiley and Sons Inc. https://doi.org/10.1002/aps3.1030

Li, W. & Godzik, A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22, 1658–1659. Available at https://doi.org/10.1093/bioinformatics/btl158

Lin, M.-Y., Koppers, N., Denton, A., Schlüter, U. & Weber, A.P.M. (2021) Whole genome sequencing and assembly data of Moricandia moricandioides and *M. arvensis*. *Data in Brief*, 35, 106922. Available at https://doi.org/10.1016/J.DIB.2021.106922

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

Mabry, M.E., Brose, J.M., Blischak, P.D., Sutherland, B., Dismukes, W.T., Bottoms, C.A. et al. (2020) Phylogeny and multiple independent whole-genome duplication events in the Brassicales. *American Journal of Botany*, 107(8), 1148–1164. Available at https://doi.org/10.1002/AJB2.1514

Mallmann, J., Heckmann, D., Bräutigam, A., Lercher, M.J., Weber, A.P., Westhoff, P. et al. (2014) The role of photorespiration during the evolution of C4 photosynthesis in the genus Flaveria. *eLife*, 3, e02478. Available at https://doi.org/10.7554/eLife.02478

Marçais, G. & Kingsford, C. (2011) A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, 27(6), 764–770. Available at https://doi.org/10.1093/BIOINFORMATICS/BTR011

McKown, A.D., Moncalvo, J.-M. & Dengler, N.G. (2005) Phylogeny of Flaveria (Asteraceae) and inference of C4 photosynthesis evolution. *American Journal of Botany*, 92(11), 1911–1928. Available at https://doi.org/10.3732/AJB.92.11.1911

Moghe, G.D., Hufnagel, D.E., Tang, H., Xiao, Y., Dworkin, I., Town, C.D. et al. (2014) Consequences of Whole-Genome triplication as revealed by comparative genomic analyses of the Wild Radish *Raphanus raphanistrum* and three other Brassicaceae species. *The Plant Cell*, 26(5), 1925–1937. Available at https://doi.org/10.1105/TPC.114.124297

Muhaidat, R., Sage, T.L., Frohlich, M.W., Dengler, N.G. & Sage, R.F. (2011) Characterization of C_3--C_4 intermediate species in the genus Heliotropium L. (Boraginaceae): anatomy, ultrastructure and enzyme activity. *Plant, Cell & Environment*, 34(10), 1723–1736.

Nagy, L.G., Merényi, Z., Hegedüs, B. & Bálint, B. (2020) Novel phylogenetic methods are needed for understanding gene function in the era of mega-scale genome sequencing. *Nucleic Acids Research*, 48(5), 2209–2219.

O'Donovan, C. (2002) High-quality protein knowledge resource: SWISS-PROT and TrEMBL. *Briefings in Bioinformatics*, 3(3), 275–284.

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*, 135(1), 15–27.

Ott, A., Schnable, J.C., Yeh, C.T., Wu, L., Liu, C., Hu, H.C. et al. (2018) Linked read technology for assembling large complex and polyploid genomes. *BMC Genomics*, 19(1), 651. Available at https://doi.org/10.1186/s12864-018-5040-z

Perfectti, F., Gómez, J.M., González-Megías, A., Abdelaziz, M. & Lorite, J. (2017) Molecular phylogeny and evolutionary history of Moricandia DC (Brassicaceae). *PeerJ*, 5, e3964. Available at https://doi.org/10.7717/PEERJ.3964/SUPP-3

Prudent, X., Parra, G., Schwede, P., Roscito, J.G. & Hiller, M. (2016) Controlling for phylogenetic relatedness and evolutionary rates improves the discovery of associations between species' phenotypic and genomic differences. *Molecular Biology and Evolution*, 33(8), 2135–2150.

Rawsthorne, S. (1992) C3–C4 intermediate photosynthesis: linking physiology to gene expression. *The Plant Journal*, 2(3), 267–274. Available at https://doi.org/10.1111/J.1365-313X.1992.00267.X

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.

Razmjoo, K., Toriyama, K., Ishii, R. & Hinata, K. (1996) Photosynthetic properties of hybrids between *Diplotaxis muralis* DC, a C3 species, and *Moricandia arvensis* (L.) DC, a C3-C4 intermediate species in Brassicaceae. *Genes & Genetic Systems*, 71(3), 189–192. Available at https://doi.org/10.1266/ggs.71.189

Reeves, G., Singh, P., Rossberg, T.A., Sogbohossou, E.O.D., Schranz, M.E. & Hibberd, J.M. (2018) Natural variation within a species for traits underpinning C_4 photosynthesis. *Plant Physiology*, 177, 504–512. Available at https://doi.org/10.1104/pp.18.00168

Roach, M.J., Schmidt, S.A. & Borneman, A.R. (2018) Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics*, 19(1), 460. Available at https://doi.org/10.1186/s12859-018-2485-7

Sage, R.F., Christin, P.-A. & Edwards, E.J. (2011) The C4 plant lineages of planet earth. *Journal of Experimental Botany*, 62(9), 3155–3169. Available at https://doi.org/10.1093/JXB/ERR048

Sage, R.F., Khoshravesh, R. & Sage, T.L. (2014) From proto-Kranz to C4 Kranz: building the bridge to C4 photosynthesis. *Journal of Experimental Botany*, 65(13), 3341–3356.

Sage, R.F., Sage, T.L. & Kocacinar, F. (2012) Photorespiration and the evolution of C4 photosynthesis. *Annual Review of Plant Biology*, 63, 19–47. Available at https://doi.org/10.1146/ANNUREV-ARPLANT-042811-105511

Schlüter, U., Bouvier, J.W., Guerreiro, R., Malisic, M., Kontny, C., Westhoff, P. et al. (2023) Brassicaceae display diverse photorespiratory carbon recapturing mechanisms. *Journal of Experimental Botany*, erad250. Available at https://doi.org/10.1093/jxb/erad250

Schlüter, U., Bräutigam, A., Gowik, U., Melzer, M., Christin, P.-A., Kurz, S. et al. (2017) Photosynthesis in C 3–C 4 intermediate Moricandia species. *Journal of Experimental Botany*, 68(2), 191–206. Available at https://doi.org/10.1093/jxb/erw391

Schlüter, U. & Weber, A.P. (2016) The road to C4 photosynthesis: evolution of a complex trait via intermediary states. *Plant Cell Physiology*, 57(5), 881–889. Available at https://doi.org/10.1093/pcp/pcw009Schuler

Schuler, M.L., Mantegazza, O. & Weber, A.P.M. (2016) Engineering C4 photosynthesis into C3 chassis in the synthetic biology age. *The Plant Journal*, 87(1), 51–65. Available at https://doi.org/10.1111/TPJ.13155

Schulze, S., Mallmann, J., Burscheidt, J., Koczor, M., Streubel, M., Bauwe, H. et al. (2013) Evolution of C4 photosynthesis in the genus flaveria: establishment of a photorespiratory CO_2 pump. *The Plant Cell*, 25(7), 2522–2535. Available at https://doi.org/10.1105/tpc.113.114520

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. Available at https://doi.org/10.1016/J.PBI.2016.03.011

Siadjeu, C., Lauterbach, M. & Kadereit, G. (2021) Insights into regulation of C2 and C4 photosynthesis in amaranthaceae/chenopodiaceae using RNA-Seq. *International Journal of Molecular Sciences*, 22(22), 12120. Available at https://doi.org/10.3390/ijms222212120

Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V. & Zdobnov, E.M. (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31(19), 3210–3212. Available at https://doi.org/10.1093/bioinformatics/btv351

Smit, A. & Hubley, R. (2015). RepeatModeler open-1.0. Available at http://www.repeatmasker.org; https://doi.org/10.1007/s00572-016-0720-5

Smit, A., Hubley, R. & Green, P. (2015) RepeatMasker Open-3.0. http://repeatmasker.org/faq.html

Smith, S.D., Pennell, M.W., Dunn, C.W. & Edwards, S.V. (2020) Phylogenetics is the new genetics (for most of biodiversity). *Trends in Ecology & Evolution*, 35(5), 415–425.

Taniguchi, Y.Y., Gowik, U., Kinoshita, Y., Kishizaki, R., Ono, N., Yokota, A. et al. (2021) Dynamic changes of genome sizes and gradual gain of cell-specific distribution of C4 enzymes during C4 evolution in genus Flaveria. *The Plant Genome*, 14(2), e20095.

The UniProt Consortium. (2019) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research*, 47(D1), D506–D515. Available at https://doi.org/10.1093/nar/gky1049

Timm, S. & Hagemann, M. (2020) Photorespiration—how is it regulated and how does it regulate overall plant metabolism? *Journal of Experimental Botany*, 71(14), 3955–3965.

Trachana, K., Larsson, T.A., Powell, S., Chen, W.H., Doerks, T., Muller, J. et al. (2011) Orthology prediction methods: a quality assessment using curated protein families. *BioEssays*, 33, 769–780. Available at https://doi.org/10.1002/bies.201100062

Trifinopoulos, J., Nguyen, L.-T., von Haeseler, A. & Minh, B.Q. (2016) W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. *Nucleic Acids Research*, 44(W1), W232–W235.

Ueno, O., Bang, S.W., Wada, Y., Kondo, A., Ishihara, K., Kaneko, Y. et al. (2003) Structural and biochemical dissection of photorespiration in hybrids differing in genome constitution between *Diplotaxis tenuifolia* (C3-C4) and radish (C3). *Plant Physiology*, 132(3), 1550–1559.

Ueno, O., Wada, Y., Wakai, M. & Bang, S.W. (2006) Evidence from photosynthetic characteristics for the hybrid origin of *Diplotaxis muralis* from a C 3 -C 4 intermediate and a C 3 species. *Plant Biology*, 8(2), 253–259. Available at https://doi.org/10.1055/s-2005-873050

Vogan, P.J. & Sage, R.F. (2012) Effects of low atmospheric CO_2 and elevated temperature during growth on the gas exchange responses of C3, C3–C4 intermediate, and C4 species from three evolutionary lineages of C4 photosynthesis. *Oecologia*, 169, 341–352.

Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S. et al. (2014) Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One*, 9(11), e112963. Available at https://doi.org/10.1371/journal.pone.0112963

Wang, C., Guo, L., Li, Y. & Wang, Z. (2012) Systematic comparison of C3 and C4 plants based on metabolic network analysis. *BMC Systems Biology*, 6(Suppl 2), S9. Available at https://doi.org/10.1186/1752-0509-6-S2-S9

Wang, O., Chin, R., Cheng, X., Wu, M.K.Y., Mao, Q., Tang, J. et al. (2019) Efficient and unique cobarcoding of second-generation sequencing reads from long DNA molecules enabling cost-effective and accurate sequencing, haplotyping, and de novo assembly. *Genome Research*, 29(5), 798–808.

Wang, P., Kelly, S., Fouracre, J.P. & Langdale, J.A. (2013) Genome-wide transcript analysis of early maize leaf development reveals gene cohorts associated with the differentiation of C4 K ranz anatomy. *The Plant Journal*, 75(4), 656–670.

Wang, P., Khoshravesh, R., Karki, S., Tapia, R., Balahadia, C.P., Bandyopadhyay, A. et al. (2017) Re-creation of a key step in the

evolutionary switch from C3 to C4 leaf anatomy. *Current Biology*, 27(21), 3278–3287. Available at https://doi.org/10.1016/J.CUB.2017.09.040

Warren, R.L., Yang, C., Vandervalk, B.P., Behsaz, B., Lagman, A., Jones, S.J.M. et al. (2015) LINKS: scalable, alignment-free scaffolding of draft genomes with long reads. *GigaScience*, 4(1), 35. Available at https://doi.org/10.1186/s13742-015-0076-3

Weber, A.P.M. & Bar-Even, A. (2019) Update: improving the efficiency of photosynthetic carbon reactions. *Plant Physiology*, 179(3), 803–812. Available at https://doi.org/10.1104/PP.18.01521

Weisenfeld, N.I., Kumar, V., Shah, P., Church, D.M. & Jaffe, D.B. (2017) Direct determination of diploid genome sequences. *Genome Research*, 27(5), 757–767. Available at https://doi.org/10.1101/gr.214874.116

Weiß, C.L., Pais, M., Cano, L.M., Kamoun, S. & Burbano, H.A. (2018) nQuire: a statistical framework for ploidy estimation using next generation sequencing. *BMC Bioinformatics*, 19(1), 122.

Wick, R.R., Judd, L.M. & Holt, K.E. (2019) Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biology*, 20(1), 129.

Xin, Z. & Chen, J. (2012) A high throughput DNA extraction method with high yield and quality. *Plant Methods*, 8(1), 26. Available at https://doi.org/10.1186/1746-4811-8-26

Yandell, M. & Ence, D. (2012) A beginner's guide to eukaryotic genome annotation. *Nature Reviews Genetics*, 13(5), 329–342.

Yeo, S., Coombe, L., Warren, R.L., Chu, J. & Birol, I. (2018) ARCS: scaffolding genome drafts with linked reads. *Bioinformatics*, 34(5), 725–731. Available at https://doi.org/10.1093/bioinformatics/btx675

Zdobnov, E.M. & Apweiler, R. (2001) InterProScan—an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, 17, 847–848. Available at https://doi.org/10.1093/bioinformatics/17.9.847

Zhang, C., Scornavacca, C., Molloy, E.K. & Mirarab, S. (2020) ASTRAL-Pro: Quartet-Based Species-Tree inference despite paralogy. *Molecular Biology and Evolution*, 37(11), 3292–3307. Available at https://doi.org/10.1093/MOLBEV/MSAA139

Zheng, G.X.Y., Lau, B.T., Schnall-Levin, M., Jarosz, M., Bell, J.M., Hindson, C.M. et al. (2016) Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nature Biotechnology*, 34, 303–311. Available at https://doi.org/10.1038/nbt.3432

SUPPORTING INFORMATION

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

How to cite this article: Guerreiro, R., Bonthala, V. S., Schlüter, U., Hoang, N. V., Triesch, S., Schranz, M. E. et al. (2023) A genomic panel for studying C3-C4 intermediate photosynthesis in the Brassiceae tribe. *Plant, Cell & Environment*, 46, 3611–3627. https://doi.org/10.1111/pce.14662

2. Brassicaceae display variation in efficiency of photorespiratory carbon-recapturing mechanisms

Here I contributed with phylogenetic work, creating Figure 2 and writing the “Phylogenetic inference” section of the Materials and Methods. I also acquired and managed seeds for the used plant accessions, germinating and growing the plants in a controlled growth chamber, as well as collecting and freezing leaf samples used for the metabolite experiment that resulted in Figures 6 and 7.

SOCIETY FOR EXPERIMENTAL BIOLOGY

Journal of Experimental Botany, Vol. 74, No. 21 pp. 6631–6649, 2023
https://doi.org/10.1093/jxb/erad250 Advance Access Publication 1 July 2023

RESEARCH PAPER

Brassicaceae display variation in efficiency of photorespiratory carbon-recapturing mechanisms

Urte Schlüter[1,*], Jacques W. Bouvier[1,†], Ricardo Guerreiro[2], Milena Malisic[1,‡], Carina Kontny[1], Philipp Westhoff[3], Benjamin Stich[2] and Andreas P.M. Weber[1]

[1] Institute of Plant Biochemistry, Cluster of Excellence for Plant Sciences (CEPLAS), Heinrich Heine University, Universitätsstr. 1, D-40225 Düsseldorf, Germany
[2] Institute for Quantitative Genetics and Genomics of Plants, Cluster of Excellence for Plant Sciences (CEPLAS), Heinrich Heine University, Universitätsstr. 1, D-40225 Düsseldorf, Germany
[3] Metabolomics and Metabolism Laboratory, Cluster of Excellence for Plant Sciences (CEPLAS), Heinrich Heine University, Universitätsstr. 1, D-40225 Düsseldorf, Germany
[†] Present address: Department of Plant Sciences, University of Oxford, South Parks Road, Oxford OX1 3RB, UK.
[‡] Present address: Department of Plant Microbe Interactions, Max Planck Institute for Plant Breeding Research, D-50829 Cologne, Germany.

[*] Correspondence: u.schlueter@hhu.de

Received 22 December 2022; Editorial decision 26 June 2023; Accepted 30 June 2023

Editor: John Lunn, MPI of Molecular Plant Physiology, Germany

Abstract

Carbon-concentrating mechanisms enhance the carboxylase efficiency of Rubisco by providing supra-atmospheric concentrations of CO_2 in its surroundings. Beside the C_4 photosynthesis pathway, carbon concentration can also be achieved by the photorespiratory glycine shuttle which requires fewer and less complex modifications. Plants displaying CO_2 compensation points between 10 ppm and 40 ppm are often considered to utilize such a photorespiratory shuttle and are termed 'C_3–C_4 intermediates'. In the present study, we perform a physiological, biochemical, and anatomical survey of a large number of *Brassicaceae* species to better understand the C_3–C_4 intermediate phenotype, including its basic components and its plasticity. Our phylogenetic analysis suggested that C_3–C_4 metabolism evolved up to five times independently in the *Brassicaceae*. The efficiency of the pathway showed considerable variation. Centripetal accumulation of organelles in the bundle sheath was consistently observed in all C_3–C_4-classified taxa, indicating a crucial role for anatomical features in CO_2-concentrating pathways. Leaf metabolite patterns were strongly influenced by the individual species, but accumulation of photorespiratory shuttle metabolites glycine and serine was generally observed. Analysis of phospho*enol*pyruvate carboxylase activities suggested that C_4-like shuttles have not evolved in the investigated *Brassicaceae*. Convergent evolution of the photorespiratory shuttle indicates that it represents a distinct photosynthesis type that is beneficial in some environments.

Keywords: *Brassicaceae*, carbon-concentrating mechanisms, C_3–C_4 intermediates, C_2 photosynthesis, photorespiration, photorespiratory glycine shuttle.

Introduction

The majority of plant species on Earth, including many crops, employ C_3 photosynthesis. In these plants, under the present environmental conditions, the central photosynthetic enzyme Rubisco fixes approximately one molecule of oxygen for every three molecules of CO_2 (Sharkey, 1988). Here, whilst the carboxylase reaction of Rubisco produces two molecules of 3-phosphglycerate (3PGA) which feeds into the Calvin–Benson–Bassham cycle (CBB), the oxygenase reaction produces 2-phosphoglycolate (2PG). 2PG is a competitive inhibitor of some CBB enzymes (Flügel *et al.*, 2017) and hence must be rapidly removed. Further, carbon contained in 2PG must be recycled into 3PGA to prevent depletion of CBB intermediates. These functions are fulfilled by the photorespiratory pathway. Photorespiration consists of coordinated enzyme activities that are located in different cellular compartments. In the plastids, 2PG is converted into glycolate followed by oxidation and transamination in the peroxisome, producing glycine. Glycine is transported into the mitochondria and metabolized into serine by glycine decarboxylation. Serine is finally converted into glycerate in the peroxisome and into 3PGA in the plastid. During glycine decarboxylation, previously fixed carbon and nitrogen are converted into CO_2 and NH_3, respectively. The re-fixation of carbon and nitrogen into organic forms, however, requires energy. Therefore, photorespiration is often considered a wasteful process in terms of energy, carbon, and nitrogen balance. Further, the affinity of Rubisco for O_2 increases with rising temperatures; in addition, stomatal closure during water shortages can lead to a drop in CO_2 concentrations inside the leaf. Thus, climate change could contribute to increased activity of the Rubisco oxygenase reaction.

Given the negative impact of photorespiration on plant productivity, there has been considerable interest in reducing flux through this pathway. For instance, various avenues to reduce photorespiratory losses which are being explored include increasing the capacity of the plant to recapture photorespiratory CO_2, modifying Rubisco kinetic properties, and introducing carbon-concentrating mechanisms to limit the oxygenase reaction of Rubisco by creating an CO_2-rich environment around the enzyme (Busch *et al.*, 2013). A better understanding of naturally occurring carbon-concentrating mechanisms will help in the design of these biotechnological approaches.

In C_3 species, photosynthesis and photorespiration mainly take place in the mesophyll (M) of the leaves. Recapture of photorespiratory CO_2 can be facilitated by arrangement of a continuous layer of plastids at the cell periphery next to the intercellular space. This arrangement acts to block diffusion of CO_2 out of the cell because the CO_2 would need to pass through the plastids where it can be reassimilated by Rubisco (Sage and Sage, 2009; Busch *et al.*, 2013). In rice, Rubisco-containing extensions (stromules) can increase the area of the plastidial barriers, preventing the efflux of CO_2 produced in the mitochondria (Sage and Sage, 2009). C_3 species such as wheat and rice achieve photorespiratory reassimilation rates of 24–38% (Busch *et al.*, 2013).

The number of chloroplasts in the C_3 bundle sheath (BS) varies between species (Leegood, 2008), but they tend to be smaller and fewer in number compared with the M. The contribution of BS chloroplasts to leaf photosynthesis is considered to be small (Kinsman and Pyke, 1998; Janacek *et al.*, 2009; Aubry *et al.*, 2014). Nevertheless, BS cells possess important roles in leaf hydraulics, phloem loading, intra-leaf signalling, and transport processes (Leegood, 2008; Aubry *et al.*, 2014; Lundgren *et al.*, 2014). Increases in BS organelle numbers indicate enhanced photosynthetic and photorespiratory activity. The centripetal arrangement of BS organelles helps to reduce loss of photorespiratory CO_2 (Muhaidat *et al.*, 2011; Sage *et al.*, 2014). Such BS cells have also been labelled as 'activated' or 'proto-Kranz' anatomy (Gowik and Westhoff, 2011; Sage *et al.*, 2014).

Further increases in BS CO_2 concentration are possible by shifting the glycine decarboxylation step from the M to the BS (Monson *et al.*, 1984; Rawsthorne *et al.*, 1988a; Sage *et al.*, 2014). This rearrangement forces photorespiratory glycine produced in the M to diffuse to the BS, where the tissue-specific increase in glycine decarboxylation activity promotes an elevated concentration of CO_2 around the BS Rubisco and thus supresses its oxygenase reaction. BS-localized activity of glycine decarboxylation is mainly associated with cell specificity of the glycine decarboxylase (GDC) P-protein (GLDP) (Rawsthorne *et al.*, 1988a; Schulze *et al.*, 2013). The installation of a glycine shuttle is accompanied by a further increase in organelle numbers in the BS. The majority of mitochondria are therefore located between the centripetally arranged plastids and the vein-orientated cell wall (Sage *et al.*, 2014). Such a combination of carbon supply by the glycine shuttle and efficient CO_2 scavenging by an adequate organelle arrangement improves the leaf carbon conservation and can be measured as reduction in the carbon compensation point (CCP or Γ). Plants employing the photorespiratory glycine shuttle are often classified as C_2 species or C_3–C_4 intermediates of type I (Edwards and Ku, 1987; Sage *et al.*, 2014). Their CO_2 reassimilation capacity was estimated to be ~73% in *Moricandia arvensis* (Hunt *et al.*, 1987).

GDC activity is thought to be absent or considerably reduced in the M cells of well-developed C_2 species, perhaps as a result of a loss-of-function mutation or insertion of a transposable element early in C_2 evolution (Rawsthorne, 1992; Sage *et al.*, 2012; Adwy *et al.*, 2015). Consistently, preferential BS localization of GLDP has been observed in many well-developed C_2 species (Lundgren, 2020; Schlüter and Weber, 2020). The assimilatory power of the BS can generally be further enhanced by decarboxylation of additional metabolites. Organelle accumulation and enhancement of organellar metabolism in the BS could increase the availability of such compounds. The glycine

shuttle not only transports carbon between the M and BS, but also creates a nitrogen imbalance between these cells where adjustment of leaf nitrogen metabolism in C_3–C_4 intermediates was proposed to occur by additional metabolite shuttling between M and BS cells (Mallmann *et al.*, 2014). Shuttling of malate, aspartate, pyruvate, α-alanine, α-ketoglutarate, and glutamate could contribute to rebalancing of nitrogen and energy balances between the two cell types (Mallmann *et al.*, 2014; Johnson *et al.*, 2021).

In the M cells, a rise in phospho*enol*pyruvate carboxylase (PEPC) activity can contribute to the provision of these shuttle metabolites. PEPC fixes carbon by catalysing the addition of bicarbonate to phospho*enol*pyruvate (PEP), forming the C_4 acid oxaloacetate that is usually quickly converted into malate or aspartate. In contrast to Rubisco, PEPC possesses higher substrate specificity and affinity. Combined with decarboxylation reactions in the BS, high PEPC activity in the M cells can implement a carbon shuttle mechanism transporting C_4 metabolites into the BS where CO_2 is released. Plants using the glycine shuttle in combination with such a C_4 shuttle have been identified mainly in the *Asteraceae* genus *Flaveria*. They are also termed C_2+C_4 or C_3–C_4 type II species (Edwards and Ku, 1987; Sage *et al.*, 2014; Bellasio, 2017).

Additional anatomic rearrangements and consequent separation of the PEPC and Rubisco reactions into the M and BS cells finally support an efficient C_4 cycle (Taniguchi *et al.*, 2021). In the M cells, CO_2 is converted into bicarbonate by carbonic anhydrase and is then fixed by PEPC. The bound carbon diffuses into the BS mainly in the form of malate or aspartate. Decarboxylation is then mediated by the NADP-malic enzyme in plastids, NAD-malic enzyme in the mitochondria, or phospho*enol*pyruvate carboxykinase in the cytosol. The cycle is completed by diffusion of a C_3 metabolite back to the M cells where ATP is needed for PEP regeneration by pyruvate phosphate dikinase. Plants with a strong C_4 shuttle, but which still exhibit Rubisco in the M, are classified as C_4-like (Moore *et al.*, 1989). In bona fide C_4 species, all CO_2 is first assimilated by PEPC, with subsequent shuttling of the resulting C_4 acid to the BS where CO_2 is then delivered to Rubisco by a decarboxylase reaction. High CO_2 partial pressure in the BS strongly represses the oxygenase reaction, the following photorespiratory pathway, and the concomitant loss of CO_2 and NH_3.

The efficiency of the C_4 shuttle also depends on anatomical features, especially the close connection between M and BS cells. In the majority of C_4 species, the BS forms a tight cell layer around the veins without direct exposure to the intercellular space (Sage *et al.*, 2014). The proportion of M tissue is reduced to a second cell layer around the BS cells, and C_4 species usually have high vein densities. Since CO_2 fixation in C_4 species can continue at lower internal CO_2 concentration (*C*i) and stomatal conductance, water use efficiency (WUE) is improved compared with C_3 species. Operation of Rubisco under high CO_2 partial pressure allows high efficiency for the carboxylase reaction with lower amounts of protein, thus also improving the nitrogen use efficiency of C_4 photosynthesis.

The continuous fitness gain in the intermediate forms seems to have been an important prerequisite for the evolution of the complex C_4 biochemistry and anatomy (Heckmann *et al.*, 2013; Williams *et al.*, 2013; Mallmann *et al.*, 2014; Blätke and Bräutigam, 2019; Dunning *et al.*, 2019a). Species using the photorespiratory glycine shuttle have been identified in various monocot and dicot plant lineages, often but not always in phylogenetic proximity to C_4 species (Sage *et al.*, 2011). Convergent evolution of the C_3–C_4 pathway indicates substantial improvement of leaf carbon economy at least under certain environmental conditions (Bellasio and Farquhar, 2019; Lundgren, 2020). Reduction in the CCP and centripetal accumulation of BS organelles seem to be general features of C_3–C_4 plants, but knowledge about the anatomical and biochemical plasticity of the pathway and their influence on leaf physiological is still limited (Schlüter and Weber, 2016).

In our study, we concentrated on the *Brassiceae* tribe that evolved ~23 million years ago in the circum-Mediterranean region (Arias and Pires, 2012). It includes multiple lineages of C_3–C_4 intermediates, but no known C_4 species (Apel *et al.*, 1997). Aiming at large sample sizes from the group of C_3 and well as C_3-C_4 species, we selected taxa from all currently known C_3–C_4 intermediates and related C_3 species from the genera *Moricandia*, *Diplotaxis*, and *Brassica*. Among the *Brassiceae* are also numerous crops species such as canola or rapeseed (*Brassica napus*), cabbage (*Brassica oleracea*), radish (*Raphanus sativus*), mustard (*Sinapis alba*), and the salad vegetable rocket or arugula (*Eruca sativa*, *Diplotaxis tenuifolia*). With the exception of *D. tenuifolia*, these are all C_3 crops.

In the present study, we analysed 34 taxa representing 28 *Brassicaceae* species. Our investigation of taxa from diverse photosynthesis types allowed us to assess the variation in 75 photosynthetic-related parameters across C_3 and C_3–C_4 species. The *Cleomaceae Gynandropsis gynadra* was included for comparison with the anatomy, biochemistry, and physiology of C_4 species. We use the CCP to rank the species and accessions according to their carbon-concentrating capacity. By parallel analysis of leaf structural features, we investigate the correlation between carbon-concentrating capacity and vein density, leaf thickness, and organelle arrangement in the BS. As the installation of the glycine shuttle requires further metabolic adjustments in the leaf such as the nitrogen balancing between M and BS cells, the primary metabolite pattern of the leaf sections was also analysed. Consequences of the anatomical and biochemical changes for the leaf physiology including assimilation and WUE were assessed. The large number of plant taxa will allow us also to learn about lineage-specific developments of the C_3–C_4 pathways and potential variation within the trait. Detailed knowledge about the carbon-concentrating mechanisms existing in *Brassiceae* can help to identify interesting traits for engineering or breeding approaches.

Materials and methods

Plant cultivation

All seeds used in this study were obtained from either Botanical gardens, seed stock centres, or seed companies. Since physiology can vary between populations from the same species (Lundgren *et al.*, 2016;Yorimitsu *et al.*, 2019; Gomez *et al.*, 2020), multiple accessions were used from some species. The complete list of plant taxa comprised: *Arabidopsis thaliana* (L.) Heynh. (At), *Brassica gravinae* Ten. four accessions, Bg1, Bg2, Bg3, and Bg4), *Brassica juncea* (L.) Czern. (Bj), *Brassica napus* L. (Bn), *Brassica nigra* (L.) W.D.J. Koch subsp. *nigra* var nigra (Bni), *Brassica oleraceae* L. (Bo), *Brassica rapa* L. (Br), *Brassica repanda* (Willd.) (Be), *Brassica tournefortii* Gouan. (two accessions, Bt1 and Bt2), *Carrichtera annua* (L.) DC. (Ca), *Diplotaxis acris* Boiss. (Da), *Diplotaxis erucoides* (L.) DC. (De), *Diplotaxis harra* Boiss. (Dh), *Diplotaxis muralis* (L.) DC. (Dm), *Diplotaxis tenuifolia* (L.) DC. (Dt), *Diplotaxis tenuisiliqua* Delile (Ds), *Diplotaxis viminea* (L.) DC. (Dv), *Eruca sativa* Mill. (Es), *Hirschfeldia incana* (L.) Lagr.-Foss (two accessions HIR1 and HIR3), *Moricandia arvensis* (L.) DC. (Ma), *Moricandia moricandioides* (Boiss.) Heywood (Mm), *Moricandia nitens* E.Durand & Barratte (Mn), *Moricandia sinaica* Boiss. (Msi), *Moricandia spinosa* Pomel (Mp), *Moricandia suffruticosa* (Desf.) Coss. & Durieu (Ms), *Raphanus raphanistrum* L. (Rr), *Raphanus sativus* subsp. *sativus* (L.) (Rs), and *Sinapis alba* L. (Sa). From the *Cleomaceae*, the C_4 species *Gymandropsis gynandra* (L.) briq. (Gg) was also included in the present study. A complete list of origins for these seeds can be found in Supplementary Table S1.

All seeds were vapour sterilized by incubation in an exicator with a fresh mixture of 100 ml of 13% Na-hypochloride with 3 ml of 37% HCl for 2 h. The sterilized seeds were then germinated on plates containing 0.22% (w/v) Murashige and Skoog medium, 50 mM MES pH 5.7, and 0.8% (w/v) agar. After 7–10 d, the seedlings were transferred to pots containing a mixture of sand and soil (Floraton 1 soil mixture, Floraguard, Germany) at a ratio of 1:2. All plants were firstly cultivated in climate chambers (CLF Mobilux Growbanks, Germany) under 12 h day conditions with 23 °C/20 °C day/night temperatures and ~200 µmol s^{-1} m^{-2} light. After establishment in soil, 2-week-old plants were transferred to the greenhouse of the Heinrich Heine University with a 16 h day/8 h night cycle. Natural light conditions in the greenhouse were supplemented with metal halide lamps (400 W, DH Licht, Germany) so that the plants received between 250 µmol m^{-2} s^{-1} and 400 µmol m^{-2} s^{-1}. Minimum temperatures of the greenhouse were controlled to 21 °C during the night and 24 °C during the day.

The initial main experiment was conducted between October 2018 and March 2019. A small number of additional accessions were also studied between July and October 2020 following the same protocol. As controls, *G. gynandra*, *D. tenuifolia*, and *H. incana* (HIR3) were included in both experiments. Gas exchange parameters obtained for these three species, especially CCPs, were stable across the experiments. Thus, results from both experiments were considered comparable. Gas exchange was measured on the youngest fully expanded rosette leaves before onset of flowering. After gas exchange measurements were performed, plants were taken back to the greenhouse for 2 d. Following that, leaf material was harvested for metabolite analysis (only experiment in 2018/2019), elemental analyser isotope ratio mass spectrometry (EA-IRMS) analysis, leaf vein determination, and embedding for light microscopy. A third experiment was conducted on plant accessions selected from the initial experiments in September to November 2021 in the new greenhouses of the Heinrich-Heine University equipped with natural light LED lamps using the same experimental design. Leaves were snap-frozen in liquid nitrogen directly in the greenhouse in the late morning Samples for protein and PEPC assay were snap-frozen in liquid nitrogen at midday. An additional leaf was used for determination of specific leaf area (SLA). For the majority of taxa, 4–6 plants were analysed as replicates in the different experiments; for a few taxa, only two plants were available for analysis. In the graphs, each data point represents a measurement from a single plant.

Phylogenetic inference

Plants for genome sequencing were grown in a climate-controlled chamber from the same seed stock used for physiological analysis. Linked read sequencing was performed by 10× Genomics and BGI, complemented by PacBio and Nanopore long read sequencing for some species (Guerreiro *et al.*, 2023). Most assemblies are linked read assemblies, some being scaffolded and gap-filled with the long read data, while two assemblies are long read assemblies polished and scaffolded by linked reads.

The assemblies are pseudohaploid, with alternative haplotype contigs having been removed with Purge Haplotigs (v1.1.0) (Roach *et al.*, 2018). The novel genome assemblies were complemented with literature assemblies (Guerreiro *et al.*, 2023). Repeat regions were identified for each assembly with Mite-hunter (Han and Wessler, 2010), genometools (v1.5.9), LTRharvest (Ellinghaus *et al.*, 2008), and RepeatModeller (v1.0.11) (Smit *et al.*, 2015). Those repeat regions were masked out of the assembly prior to gene annotation using RepeatMasker (v4.0.9) (Smit *et al.*, 2015). Gene annotation was performed using Maker2 (Campbell *et al.*, 2014) and the same protein database for every genome (Guerreiro *et al.*, 2023). The predicted proteomes of every species were filtered for annotation edit distance (AED; Eilbeck *et al.*, 2009) values <0.5. Functional annotation was performed with AHRD in order to remove transposon-related genes (Guerreiro *et al.*, 2023).

Finally, the filtered proteomes were fed into Orthofinder v2.5.1 (Emms and Kelly, 2015, 2019) for orthology identification based on all versus all sequence BLASTp searches and MCL clustering (Emms and Kelly, 2015). Multiple sequence alignments for identified orthogroups (HOGs) were produced with MAFFT and used for creating gene trees with RaxML with the PROTGAMMALG substitution model. The gene trees of HOGs with single-copy genes for at least 80% of species (102 HOGs) were fed to ASTRAL-pro (Zhang *et al.*, 2020) for the creation of a species phylogeny with quartet-based local posterior probability values (Sayyari and Mirarab, 2016) for each node.

Photosynthetic gas exchange

Gas exchange was measured on the youngest fully expanded rosette leaf ~6–10 weeks after sowing and before the onset of flowering. The settings of the LI-6800 (LI-COR, Lincoln, NE, USA) were as follows: flow of 300 µmol s^{-1}, fan speed of 10 0000 rpm, light intensity of 1500 µmol m^{-2} s^{-1}, leaf temperature of 25 °C, and vapour pressure deficit of 1.5 kPa. After adjustment of leaves to the conditions in the leaf chamber, *A*–*Ci* curves were measured at reference atmospheric CO_2 concentrations of 400, 200, 100, 75, 50, 40, 30, 20, 10, 0, 400, 400, 600, 800, 1200, and 1600 ppm. For the experiment in 2018/2019, the LI-6800 was equipped with a fluorescence head measuring F_v'/F_m' and electron transport rate (ETR) at each CO_2 level.

For calculation of the CCP and the carboxylation efficiency (CE=initial slope of *A*–*Ci*), a minimum of four data points in the linear range close to the interception with the *Ci* axis were used. Maximal assimilation was determined at CO_2 concentrations between 1200 ppm and 1600 ppm. Assimilation (*A*), stomatal conductance (g_{sw}), *Ci*, WUE=A/g_{sw}, and the ratio between internal and external CO_2 concentrations (*Ci*/*Ca*) from the measurements at 400 (ambient CO_2), 200, 100, and 50 ppm CO_2 were used for more detailed physiological analysis of the investigated plant accessions.

Metabolite and element analysis

After the gas exchange measurements, plants were allowed to readjust to greenhouse conditions before sampling for metabolite patterns. Leaves were snap-frozen in liquid nitrogen directly in the greenhouse in the late morning and stored at −80 °C. The leaf samples then were

homogenized into a fine powder by grinding in liquid nitrogen. Soluble metabolites were extracted in 1.5 ml of extraction solution consisting of water:methanol:chloroform in a 1:2.5:1 mixture following the method of Fiehn *et al.* (2000). A 30 µl aliquot of the supernatant was dried completely in a vacuum concentrator and derivatized for GC-MS measurements (Gu, 2012). GC-MS measurements were performed as described by Shim *et al.* (2020) using a 5977B GC-MSD (both Agilent Technologies). Metabolites were identified via MassHunter Qualitative (v b08.00, Agilent Technologies) by comparison of spectra with the NIST14 Mass Spectral Library (https://www.nist.gov/srd/nist-standard-reference-database-1a-v14). A standard mixture containing all target compounds at a concentration of 5 µM was processed in parallel to the samples as a response check and retention time reference. Peaks were integrated using MassHunter Quantitative (v b08.00, Agilent Technologies). For relative quantification, all metabolite peak areas were normalized to the corresponding fresh weight used for extraction and to the peak area of the internal standard ribitol or dimethylphenylalanine (Sigma-Aldrich). The same homogenized leaf material was used for determination of $\delta^{13}C$ and CN ratios. After lyophilization, the material was analysed using an Isoprime 100 isotope ratio mass spectrometer coupled to an ISOTOPE cube elemental analyser (both from Elementar, Hanau, Germany) according to Gowik *et al.* (2011).

Vein density measurements

The top third of mature rosette leaves were used for vein density measurements. The leaf material was cleared in an acetic acid:ethanol mix (1:3) overnight followed by staining of cell walls in 5% safranin O in ethanol, and de-staining in 70% ethanol. Pictures were taken using a Nikon eclipse Ti-U microscope equipped with a ProgRes MF camera from Jenoptik, Germany, at ×4 magnification. The vein density was analysed with ImageJ software determining the total vein length per total micrograph area. In most cases, six leaves were analysed for vein density per line with a minimum of three pictures measured and averaged per leaf.

Specific leaf area

Whole mature rosette leaves were cut and their outlines were copied on checked paper. The FW was measured immediately after and DW was subsequently determined after 48 h at 60 °C. The leaf area was determined using ImageJ software. For calculation of SLA, the area was divided by the dry weight.

Analysis of leaf cross-section

For light microscopy, sections of ~1 × 2 mm were cut from the top third of mature rosette leaves and immediately fixed in 2% paraformaldehyde, 2% glutaraldehyde, 0.1% Triton X-100 in phosphate-buffered saline (PBS; 137 mM NaCl, 2.7 mM KCl, 12 mM $H_2PO_4^-/HPO_4^{2-}$, pH 7.4). A vacuum was applied to the reaction tubes until all leaf sections sank to the bottom. The sections were incubated in the primary fixation solution overnight followed by washing once with PBS solution, pH 7.4, and twice with distilled H_2O. For post-fixation, the sections were incubated in 1% OsO_4 for 45 min followed by washing again three times with distilled H_2O. A dehydration series was performed ranging from 30% to 100% acetone, followed by incubation in increasing proportions of Araldite resin until 100% Araldite was reached. The sections were finally positioned into flat embedding moulds and polymerized at 65 °C for at 48 h.

After cutting, the leaf sections were stained with toluidine blue solution (0.5% toluidine blue, 0.5% methylene blue, 6% $Na_2B_4O_7$, 1% H_3BO_3) and studied under a light microscope (Zeiss Axio Observer, Carl Zeiss, Germany). For quantitative analysis, pictures of at least three BSs per biological replicate were taken and analysed with ImageJ software. The following parameters were determined per BS for quantitative analysis: cross-section area of the BS cell (BS_cell_area), area of organelles orientated towards the vein (V_organelle_area), area of organelles orientated towards the intercellular space (ICS), and M (M_organelle_area). The following parameter were calculated: percentage of vein-orientated organelle area per BS cell area (percent_V_organelle), percentage of organelle area orientated towards the ICS and M (percent_M_organelles), the total organelle area per cell (V_organelle_area + M_organelle_area = total_organelle_area), and the ratio between percentage values for vein and ICS/M-orientated organelles. Furthermore, the leaf thickness was measured at the site of the selected BS. Three representative cells were analysed per BS, and three BSs were analysed per plant.

PEPC activity

Total soluble proteins were extracted from the homogenized leaf material in 50 mM HEPES-KOH pH 7.5, 5 mM $MgCl_2$, 2 mM DTT, 1 mM EDTA, 0.5% Triton X-100. For the PEPC assay, 10 µl of the extract were mixed with assay buffer consisting of 100 mM Tricine-KOH pH 8.0, 5 mM $MgCl_2$, 2 mM DTT, 1 mM $KHCO_3$, 0.2 mM NADH, 5 mM glucose-6-phosphate, and 2 U ml^{-1} malate dehydrogenase in a microtitre plate. The reaction was started after addition of PEP to a final concentration of 5 mM in the assay. Protein content of the solutions was determined with the BCA assay (Thermo Fischer Scientific).

Statistical analysis

Data analysis was performed using R (www.R-project.org). Statistical differences between the measured parameters in the accessions were calculated by one-way ANOVA followed by Tukey's post-hoc test. Differences between parameters in C_3 and C_3–C_4 photosynthesis types were determined by a two-tailed *t*-test.

Results

Assessment of CO_2-concentrating efficiency by measuring CO_2 compensation points

The CCP is a measure of the internal leaf CO_2 level at which photosynthetic CO_2 fixation is equal to the CO_2 release by photorespiration, day respiration, and other catalytic processes (i.e. the concentration at which net CO_2 assimilation is zero). In the present study, the CCP was determined from *A*–*C*i curves across a diverse range of 33 *Brassicaceae* taxa representing 28 species to assess their carbon usage efficiency (Fig. 1).

When sorting all sampled taxa in the *Brassicaceae* according to their CCP, a range of CCP values from 60 ppm to 12 ppm was detected. The *Cleomaceae* C_4 species *G. gynandra* with highly efficient CO_2 concentration showed CCP values <10 ppm. Plants with CCP values between 10 ppm and 40 ppm are predicted to utilize less efficient CO_2-concentrating mechanisms, such as the photorespiratory shuttle, and were hereby classified as C_3–C_4 intermediates. In contrast, all accessions and species with a CCP >40 ppm were classified as C_3 species. This grouping was supported by ANOVA with post-hoc Tukey's HSD test (α=0.05; Fig. 1). The same threshold values were also used in the survey by Krenzer *et al.* (1975).

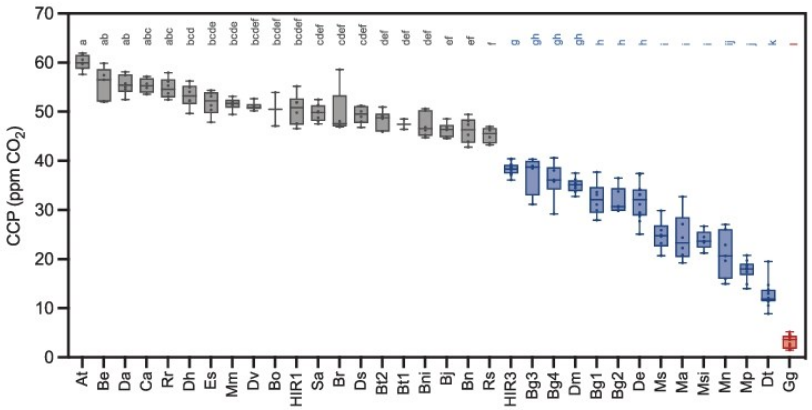

Fig. 1. CO_2 compensation points in selected *Brassicaceae*. CO_2 compensation points were measured in young, fully expanded leaves of greenhouse-grown plants. The letters above each box indicate the statistical grouping determined by ANOVA followed by HSD post-hoc test with α=0.05. The tested taxa are coloured according to photosynthesis type as C_3 (grey), C_3–C_4 (blue), and C_4 (red). Plant names have been abbreviated for legibility and are provided in full in the Materials and methods.

In our study, *A. thaliana* exhibited the highest CCP of 60.1 ppm. Slightly lower CCPs between 45 ppm and 60 ppm were observed for the C_3 species of the *Brassiceae* clade, including *B. repanda*, *D. acris*, *C. annua*, *R. raphanistrum*, *D. harra*, *E. sativa*, *M. moricandioides*, *D. viminea*, *B. oleraceae*, *H. incana* HIR1, *S. alba*, *B. rapa*, *D. tenuisiliqua*, *B. tournefortii*, *B. nigra*, *B. juncea*, *B. napus*, and *R. sativus*. In comparison with these C_3 plants, a significant reduction in the CCPs was observed in 14 taxa classified in the present study as C_3–C_4 intermediates, these included *H. incana* HIR3, four taxa of *B. gravinae*, *D. muralis*, *D. erucoides*, *M. suffruticosa*, *M. arvensis*, *M. sinaica*, *M. nitens*, *M. spinosa*, and *D. tenuifolia*.

Among the C_3–C_4 intermediates, the lowest CCP value of 12 ppm was measured in *D. tenuifolia*. In contrast, the highest CCP value recorded among the C_3–C_4 intermediates at ~40 ppm was measured in *H. incana* HIR3. Importantly, the identification of this taxon as a C_3–C_4 intermediate species which presumably operates a CO_2-concentrating mechanism is described here for the first time. Interestingly, another taxon assigned to the same species (*H. incana* HIR1) exhibited a CCP value within the range of C_3 species (Fig. 1). In all other species, different taxa were assigned to the same photosynthetic type. Altogether, a wide range of CCPs was observed among *Brassicaceae* and especially the C_3–C_4 intermediates.

Phylogeny suggests multiple origins of C_3–C_4 photosynthesis in the Brassiceae

The phylogenetic relationship among the plant species selected for this study was investigated using sequence data from 102 orthogroups (Fig. 2). When investigating the distribution of species classified as C_3–C_4 intermediates based on CCP data, the tree reveals multiple origins of CO_2-concentrating mechanisms in the *Brassiceae*. We are aware that the presented tree includes only a small subset of species from the *Brassiceae* group. However, a more densely sampled phylogenetic tree by Koch and Lemmel (2019) suggests the same number of origins because our predicted C_3–C_4 lineages all have common ancestors with C_3 species.

The phylogenetic positions of *B. gravinae*, *H. incana* HIR3, and *D. erucoides* on the tree suggest independent origins of C_3–C_4 features. In the *Moricandia* group, C_3–C_4 features are observed among the close relatives *M. arvensis*, *M. suffruticosa*, *M. nitens*, *M. sinaica*, and *M. spinosa*, but not the sister species *M. moricandioides* which is C_3. Thus, C_3–C_4 most probably evolved once in this group, in a common ancestor preceding the speciation of *M. arvensis*, *M. suffruticosa*, *M. nitens*, *M. sinaica*, and *M. spinosa* but after diversification from *M. moricandioides*. Finally, C_3–C_4-like CCPs were observed in *D. tenuifolia* and *D. muralis*. Both of these respective species are closely related to the C_3 species *D. viminea*. *Diplotaxis muralis* is a natural hybrid between the C_3 parent *D. viminea* and the C_3–C_4 parent *D. tenuifolia*. Therefore, the C_3–C_4 features in *D. muralis* are assumed to be inherited from *D. tenuifolia* (Ueno *et al.*, 2006). In summary, our phylogenetic data indicate that C_3–C_4 features have independently evolved up to five times in the *Brassicaceae*, in *B. gravinae*, in *D. erucoides*, in *H. incana* HIR3, in the *Moricandia* group, and in *D. tenuifolia*.

Physiology of C_3, C_3–C_4, and C_4 leaves under different CO_2 concentrations

Efficiency of photosynthetic gas exchange in the different C_3 and C_3–C_4 *Brassicaceae* was assessed under ambient CO_2 (400 ppm) and saturating light (1500 µmol m^{-2} s^{-1}). Here, under these conditions, no association between photosynthesis

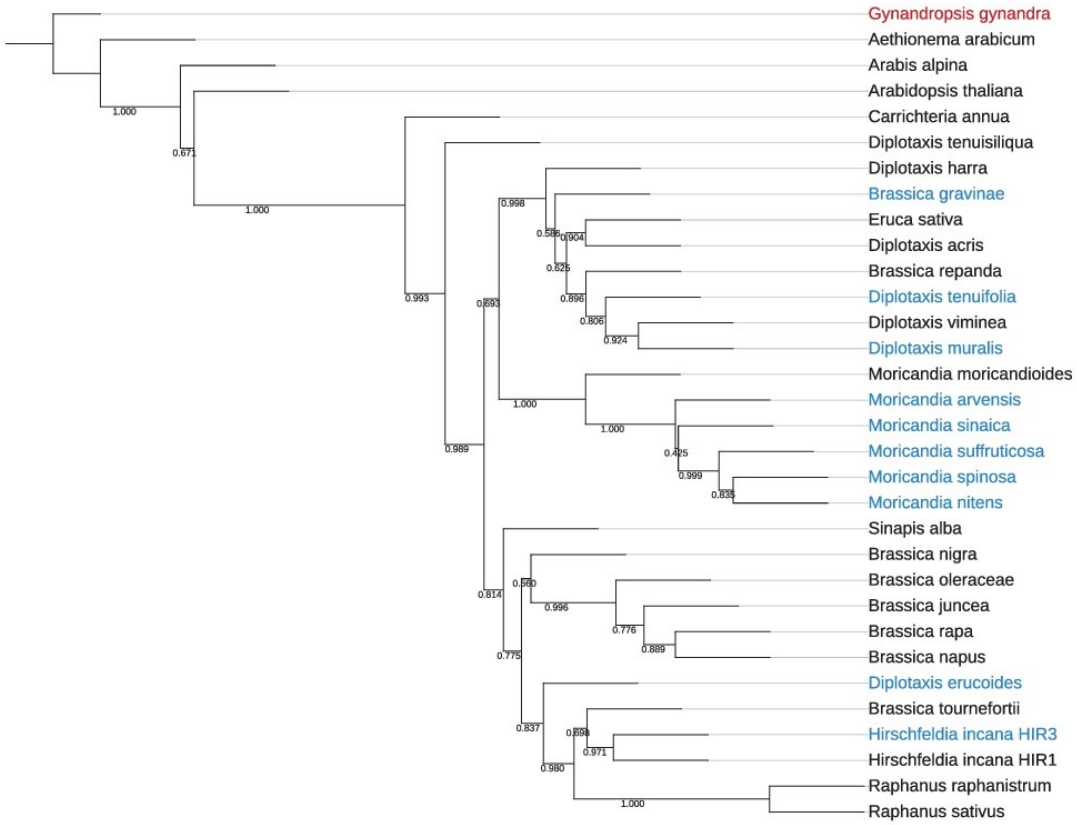

Fig. 2. Phylogeny and photosynthesis types. The numbers on the nodes represents quartet-based local posterior probability values. Species names are coloured according to the photosynthesis type as C_3 (grey), C_3–C_4 (blue), and C_4 (red).

type and net assimilation could be observed (Fig. 3). For instance, in C_3 plants, assimilation rates under ambient CO_2 varied between 12.3 μmol m^{-2} s^{-1} (*A. thaliana*) and 28.1 μmol m^{-2} s^{-1} (*D. tenuisiliqua*) (Fig. 3; Supplementary Table S2), whilst among C_3–C_4 intermediates, assimilation rates varied between 17.3 μmol m^{-2} s^{-1} (*B. gravinae* accession 2) and 26.1 μmol m^{-2} s^{-1} (*D. erucoides*). Moreover, assimilation rates achieved in the C_4 species *G. gynandra* of 23.7 μmol m^{-2} s^{-1} were similar to rates in non-C_4 plants. Thus, assimilatory capacity under ambient CO_2 conditions appears to be species specific, rather than determined by the activity of a metabolite shuttle mechanism.

In contrast to the above, enhanced rates of assimilation were discovered in plants operating a CO_2-concentrating mechanism under lower atmospheric CO_2 concentrations (Fig. 3; Supplementary Table S2). For instance, at pre-industrial levels of 200 ppm CO_2, the C_4 *G. gynandra* showed higher assimilatory capacity compared with any other C_3 or C_3–C_4 species (Fig. 3; Supplementary Table S2). Further, this elevated assimilation rate observed in *G. gynandra* became even more pronounced under 100 ppm CO_2 (Fig. 3; Supplementary Table S2). Interestingly, C_3–C_4 intermediate species tended to perform better than C_3 plants under subambient CO_2 conditions (Fig. 3; Supplementary Table S2). On average, across all plants of each photosynthesis type, assimilation was significantly higher in the C_3–C_4 group compared with the C_3 group under CO_2 conditions of ≤200 ppm (*t*-test, $P<0.05$; Supplementary Fig. S1). Thus, although CO_2-concentrating mechanisms do not improve net assimilation under present atmospheric CO_2, they appear to be advantageous under former pre-industrial levels of CO_2.

In addition to the above, operating a CO_2-concentrating mechanism also yielded benefits in terms of improved WUE (ratio between CO_2 assimilation and water stomatal conductance). For example, under ambient 400 ppm CO_2, the WUE was significantly higher in the C_4 species *G. gynandra* as compared with all other C_3 and C_3–C_4 species (Fig. 4). On average, WUE did not differ between C_3 and C_3–C_4 accessions at 400 ppm CO_2. However, C_3–C_4 plants were found to exhibit a significantly improved WUE at both 200 ppm and 100 ppm CO_2 compared with C_3 species (*t*-test, $P>0.05$; Supplementary

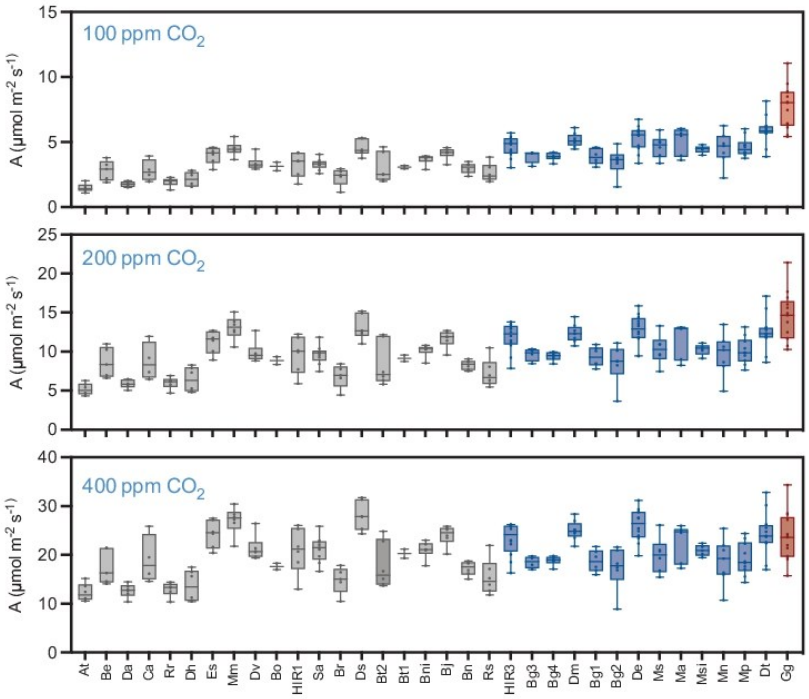

Fig. 3. Net assimilations under different CO_2 concentrations. Assimilation was measured under conditions of ambient CO_2 (400 ppm) or reduced CO_2 concentrations of 200 ppm and 100 ppm. The tested taxa were sorted according to their CO_2 compensation points and coloured according to the photosynthesis type as C_3 (grey), C_3–C_4 (blue), and C_4 (red). Taxa names have been abbreviated for legibility and are provided in Fig. 2 and the Materials and methods.

Fig. S1), recapitulating the trend observed for the assimilation rate. In addition, a strong negative correlation between WUE and Ci was found across species at all atmospheric CO_2 concentrations (Supplementary Fig. S2). Given that changes in stomatal conductance exhibited no photosynthesis type-specific pattern (Supplementary Fig. S1), this result suggests that higher WUE is achieved across species in the *Brassicaceae* by CO_2 assimilation at lower internal Ci, Thus, C_3–C_4 species are able to operate at lower internal CO_2 concentrations than species from the C_3 group.

Next, to determine the effect of C_3, C_3–C_4, and C_4 metabolism on the light reactions of photosynthesis, chlorophyll fluorescence parameters were also measured. In general, a positive correlation was found between net assimilation and ETR, as well as between assimilation rate and effective quantum efficiency F_v'/F_m' (Supplementary Fig. S2). However, fluorescence parameters were less affected under reduced atmospheric CO_2 concentrations as compared with the assimilation rate (Supplementary Fig. S2).

To further assess how different photosynthesis types are characterized by differences in leaf physiological parameters, a principal component analysis (PCA) was performed. Since the importance of carbon-concentrating mechanisms becomes more obvious when CO_2 is limited, gas exchange measurements at CO_2 concentrations of 200, 100, and 50 ppm were included in this analysis in addition to those measured under ambient 400 ppm CO_2 (Fig. 5A, B). In this PCA, the first principal component was found to explain 65.5% of the variation, and separates the C_4 species *G. gynandra* from all other C_3 and C_3–C_4 plants. To a lesser extent, the same component also separates C_3 and C_3–C_4 plants, though these groups do overlap on this axis (Fig. 5A, B). As expected from the above results, the first principal component is driven by WUE, Ci, Ci/Ca ratio, CCP, and CE. In contrast, the second principal component, driven by stomatal conductance and assimilation at higher CO_2 concentrations, has no effect on separating plants across different photosynthesis types and is driven by species-specific variation.

As CCP was used above to classify different photosynthesis types, correlations were investigated between CCP values across species and all other measured leaf physiological parameters. A strong positive correlation was observed between CCP and both Ci and the Ci/Ca ratio at low CO_2 concentrations, respectively (Fig. 5C, D). In contrast, a negative correlation was found between CCP and both WUE and assimilation at low CO_2 (Fig. 5C, D). These relationships were independently

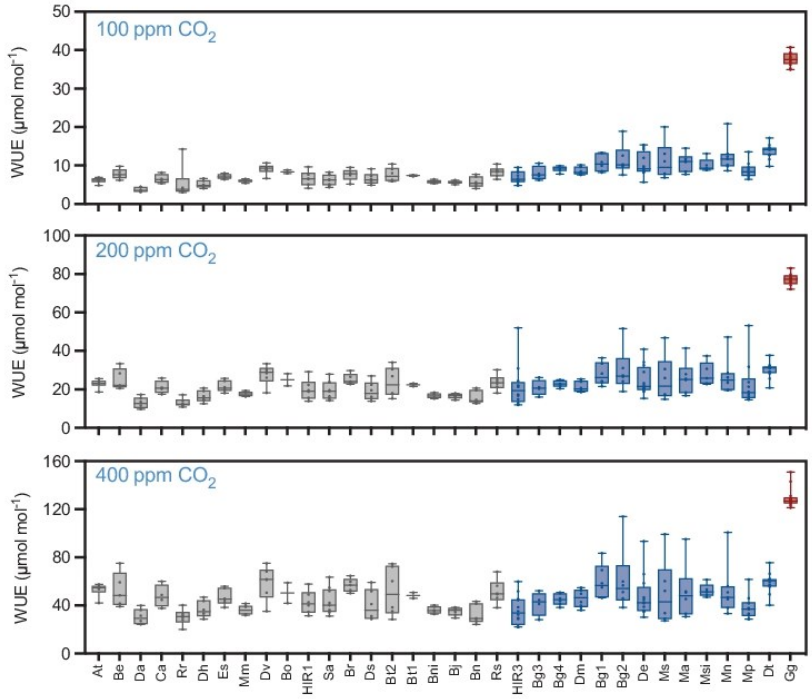

Fig. 4. Water use efficiency (WUE) under different CO_2 concentrations. WUE was calculated as the ratio between assimilation (*A*) and stomatal conductance (g_{sw}). The gas exchange parameters were measured under conditions of ambient CO_2 (400 ppm) or reduced CO_2 concentrations of 200 ppm and 100 ppm. The tested taxa were sorted according to their CO_2 compensation points and coloured according to the photosynthesis type as C_3 (grey), C_3–C_4 (blue), and C_4 (red). Taxa names have been abbreviated for legibility and are provided in Fig. 2 and the Materials and methods.

observed irrespective of whether the C_4 species *G. gynandra* was included in the analysis or not. Conversely, CE (=initial slope of the *A*–*Ci* curve) was negatively correlated with CCP only when the C_4 *G. gynandra* was included in the analysis (Fig. 5C, D). This indicates that CE was not influenced considerably during the transition from C_3 to C_3–C_4, but only during transition from C_3–C_4 to C_4.

Metabolite profiles of leaves with different photosynthesis pathways

To assess the identity of potential transport metabolites used in C_3–C_4 intermediates, leaf primary metabolites of sampled species were also quantified and analysed by PCA (Fig. 6A, B). In this analysis, the first principal component explained 28.25% of variation and distinguishes C_3/C_3–C_4 and C_4 leaf biochemistry. Mainly responsible for this separation are high levels of α-alanine, α-ketoglutarate, aspartate, glycine, glutamate, pyruvate, phenylalanine, and γ-aminobutyric acid (GABA) in the C_4 *G. gynandra* compared with the C_3 and C_3–C_4 background (Fig. 6B; Supplementary Table S3). In contrast, the second principal component sorts the majority of C_3 species (clustered to the top of this axis) from C_3–C_4 species (clustered to the bottom of this axis) (Fig. 6A, B). Here, on this axis, the C_3–C_4 plants tend to have higher levels of serine, branched amino acids, and proline, whilst the C_3 species are characterized by higher levels of glucose, sucrose, and myo-inositol.

Correlation analyses of CCP values with primary metabolite levels were also performed across species (Fig. 6C, D; Supplementary Fig. S3). A negative correlation was observed between CCP and the C_4-related metabolites α-alanine, α-ketoglutarate, aspartate, glutamate, and pyruvate (Figs 6, 7; Supplementary Table S3). However, this relationship seems to be driven by the strong accumulation of these metabolites in the C_4 species alone (Fig. 6C). To identify metabolites that are potentially specific to C_3–C_4 photosynthesis, the correlation analysis was repeated without the C_4 outgroup species. This resulted in the reduction of the strength of all statistical associations. Specifically, only serine and glycine showed significant negative correlations with CCP (Fig. 6D; Supplementary Table S3). Thus, this suggests that serine and glycine have a ubiquitous role in the glycine shuttle across all C_3–C_4 intermediates in the *Brassiceae*. Interestingly, however, glycine was the only metabolite that increased between C_3 and C_3–C_4 species which

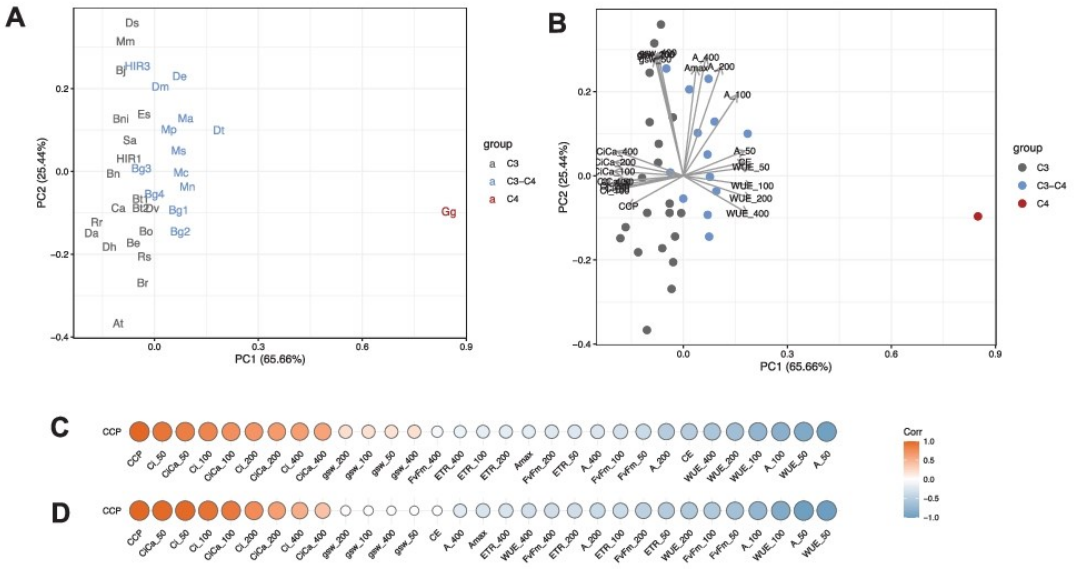

Fig. 5. Principal component analysis (PCA) and correlations of the CO_2 compensation point (CCP) with selected gas exchange parameters. Average values for the selected photosynthetic parameters determined under 50–400 ppm of CO_2 were used for the analysis. (A) Localization of the plant lines in the PCA, (B) PCA including parameter loadings, (C) Pearson correlation coefficients demonstrated as heatmaps using all taxa, and (D) Pearson correlation coefficients demonstrated as heatmaps using only C_3 and C_3–C_4 lines. The tested taxa were coloured according to photosynthesis types as C_3 (grey), C_3–C_4 (blue), and C_4 (red). Taxa names have been abbreviated for legibility and are provided in Fig. 2 and the Materials and methods. The dataset included CO_2 compensation point (CCP), carboxylation efficiency (CE), assimilation (A), internal CO_2 concentration (Ci), stomatal conductance (gsw), water use efficiency (WUE), ratio of internal to external CO_2 concentrations (CiCa), electron transport rate (ETR), and quantum efficiency F_v'/F_m' (FvFm). The numbers after the parameter abbreviation indicate the CO_2 concentration in the outside the leaf in the measuring cuvette.

was also high in the C_4 species (Fig. 7A; Supplementary Figs S3, S4). In contrast, serine was enhanced among C_3–C_4 species compared with C_3 species, but was detected at a C_3 level in the C_4 *G. gynandra* (Fig. 7B; Supplementary Figs S3, S4). In the present study, glutamate, α-alanine, aspartate, pyruvate, malate, and α-ketoglutarate formerly predicted to be involved in nitrogen shuttling of C_3–C_4 leaves (Mallmann *et al.*, 2014) were not associated with CCP (Figs 6D, 7). Instead, levels of these metabolites were high in only some, but not all, C_3–C_4 taxa. For instance, glutamate and aspartate levels were relatively high in *M. arvensis*, *D. muralis*, and *D. tenuifolia*, but not in the other C_3–C_4 *Moricandia* species *M. nitens* and *M. suffruticosa* (Fig. 7). In contrast, *D. erucoides* separated from the majority of other C_3–C_4 species in the PCA (Fig. 6A, B), showing relatively high levels of glycerate, glycolate, and malate (Fig. 7; Supplementary Table S3). In summary, the present results describe a general role for only glycine and serine as predicted shuttle metabolites in C_3–C_4 biochemistry across all species.

Association of CCP with structural features of the bundle sheath

Given that leaf and BS cell architecture play an important role in underpinning CO_2-concentrating mechanisms by enabling adequate metabolite transport between M and BS tissue, we also sought to characterize the leaf anatomy of our *Brassicaceae* species. In the present study, it was observed that vein density was highest in the C_4 *G. gynandra* compared with all other species. However, no difference in vein density was observed between C_3 and C_3–C_4 plant accessions (Fig. 8A; Supplementary Fig. S5). To determine whether differences in BS structure were present between photosynthetic types, a representative subset of plant accessions were studied in more detail by light microscopy (Supplementary Fig. S6). In this analysis, although BS cross-section area was high in the C_4 species as well as in several C_3–C_4 species, it was not found to be significantly different between C_3 and C_3–C_4 plants (Fig. 8B; Supplementary Fig. S7). Moreover, within the BS cells, the areas occupied by plastids and other organelles with either vein (inner half) or ICS/M orientation (outer half) were determined. Areas with ICS/M-oriented organelles did not differ between C_3 and C_3–C_4 leaf cross-sections (Fig. 8E). In the C_4 leaf, none of the BS organelles faced the outer ICS/M side. On the other hand, all plant accessions with a CCP <40 ppm featured enhanced organelle accumulation around the vein (Fig. 8H). This resulted in higher total organelle area in the C_3–C_4 BS cells. Thus, organelle abundance and orientation probably played a decisive role for the functioning of weak CO_2-concentrating mechanisms.

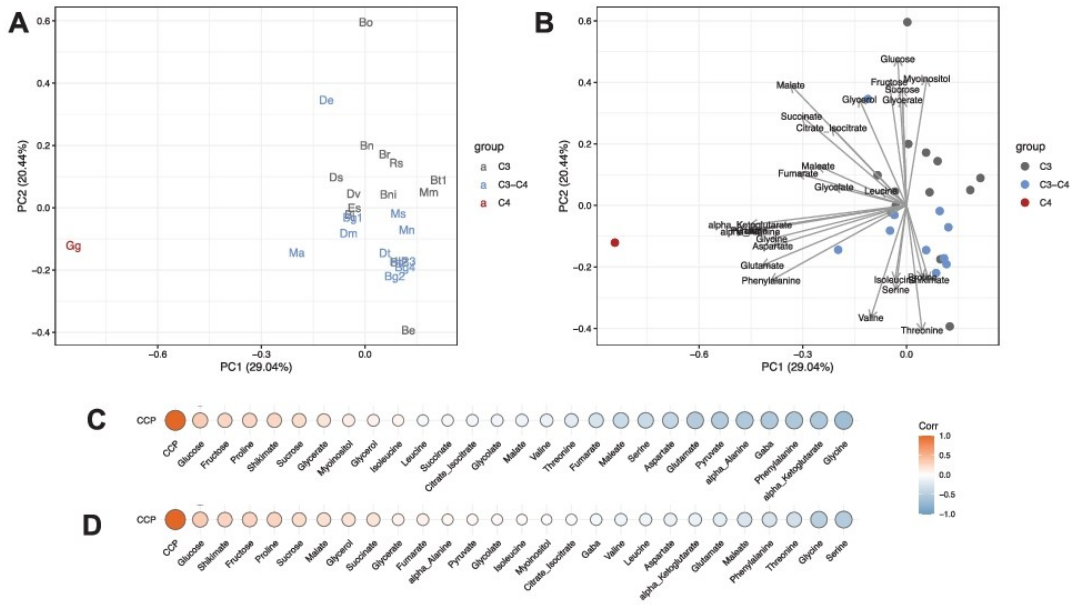

Fig. 6. Principal component analysis (PCA) and correlations of the CO_2 compensation point (CCP) with specific metabolites. Average values for the selected metabolites per taxon were used for the analysis. (A) Localization of the plant lines in the PCA, (B) PCA including metabolite loadings, (C) Pearson correlation coefficients demonstrated as heatmaps using all plant lines, and (D) Pearson correlation coefficients demonstrated as heatmaps using only C_3 and C_3–C_4 lines. The tested taxa were coloured according to photosynthesis type as C_3 (grey), C_3–C_4 (blue), and C_4 (red). Taxa names have been abbreviated for legibility and are provided in Fig. 2 and the Materials and methods.

C_4 anatomy consists of just one layer of BS and M cells around the veins, which limits the total number of cell layers. In our study, the C_4 leaves of *G. gynandra* were comparably thin. However, leaf thickness within the C_3 and C_3–C_4 groups showed species-specific variation. For instance, independent of photosynthesis type, all *Moricandia* species possessed thick succulent leaves. Values of SLA are usually greater for C_4 than for C_3 leaves (Atkinson *et al.*, 2016), but no pronounced photosynthesis type-related differences in SLA could be observed in our study (Fig. 8I).

The C_4 pathway allows plants to fix CO_2 with lower nitrogen input. This means that typically, C_4 plants have lower leaf nitrogen concentrations compared with C_3 species (Long, 1999; Craine *et al.*, 2005; Gowik *et al.*, 2011). Interestingly, however, C_4 *G. gynandra* in our analysis had surprisingly high leaf nitrogen concentrations and low leaf CN ratios (Fig. 8D). This result could be influenced by the slow growth rate of this species in comparison with the majority of *Brassicaceae* species in this study. However, leaf protein concentrations in this C_4 *G. gynandra* were low relative to the background of other species, indicating C_4-specific differences in nitrogen allocation (Fig. 8F). Interestingly, no difference in CN ratios or leaf protein concentrations could be observed between the C_3 and C_3–C_4 species (Fig. 8D; Supplementary Fig. S7).

Operation of the C_4 pathway required increased activity of PEPC, but allows reduction in concentrations of Rubisco and CBB cycle enzymes (Brautigam *et al.*, 2011; Gowik *et al.*, 2011). In our study, PEPC activity was 8- to 20-fold higher in the C_4 *G. gynandra* leaves as compared with the leaves of the C_3 and C_3–C_4 species (Fig. 8C; Supplementary Table S4). PEPC activities varied in the individual plant taxa, but were not significantly different between the C_3 and C_3–C_4 groups (Supplementary Fig. S7). *Enuca sativa* and *M. moricandioides*, in particular, showed PEPC activities similar to the C_3–C_4 taxa *M. arvensis*, *M. suffruticosa*, and *D. tenuifolia* (Fig. 8; Supplementary Fig. S7). These results emphasize the power of our multispecies analysis that allows distinction between species- and photosynthesis type-related variation.

Summarizing the above-mentioned structural and leaf composition-related parameters in a PCA, the C_4 *G. gynandra* can be separated from the rest of the *Brassicaceae* plants (Fig. 9A, B). This was mainly driven by high values for $\delta^{13}C$, vein density, and vein-orientated organelles in the BS as well as low values for CCP and ICS/M-orientated organelles in the BS (Fig. 9A, B). C_3 and C_3–C_4 accessions separated along the same line in a combination of PC1 and PC2, but an overlap between the two groups was nevertheless observed. Correlation of the CCP to the selected components supported the importance of organelle accumulation and orientation in the BS for the

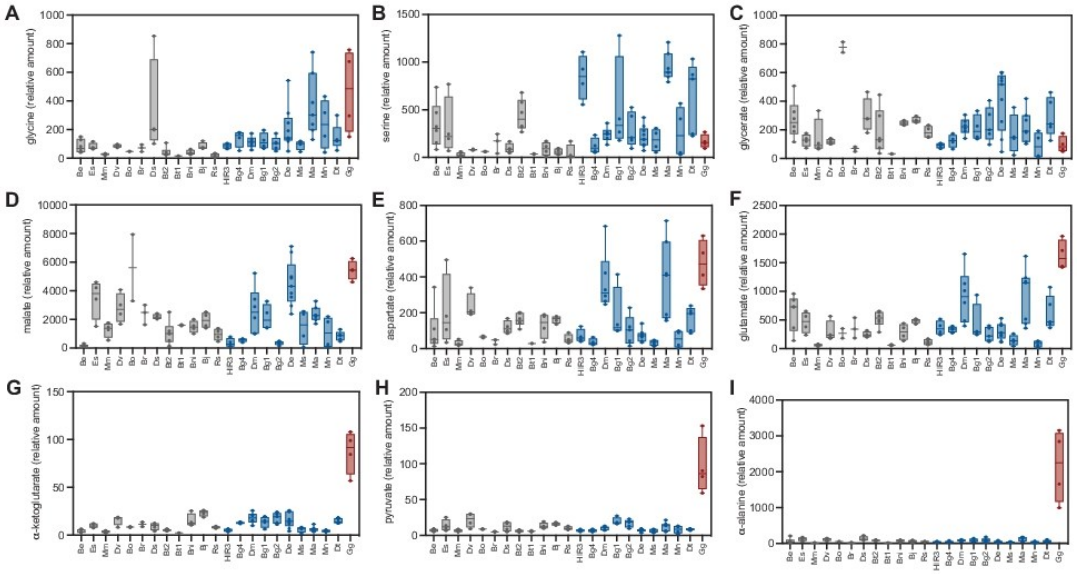

Fig. 7. Metabolites in mature leaves of selected *Brassicaceae*. Relative amounts of glycine (A), serine (B), glycerate (C), malate (D), aspartate (E), glutamate (F), α-ketoglutarate (G),and pyruvate and α-alanine in selected plant accessions. The tested taxa were sorted according to their CO_2 compensation points and coloured according to photosynthesis type as C_3 (grey), C_3–C_4 (blue), and C_4 (red). Taxa names have been abbreviated for legibility and are provided in Fig. 2 and the Materials and methods.

activity of the C_4 as well as the C_3–C_4 pathway (Fig. 9C, D; Supplementary Fig. S8).

Discussion

Physiological and phylogenetic analysis indicate evolution of multiple independent C_3–C_4 lineages in the Brassiceae

Our survey revealed multiple origins of C_3–C_4 photosynthesis in the *Brassiceae* tribe (Figs 1, 2), ranging from very efficient photorespiratory shuttles in *D. tenuifolia* and the *Moricandia* genus (*M. suffruticosa*, *M. arvensis*, *M. sinaica*, *M. nitens*, and *M. spinosa*), to relatively weaker carbon-concentrating mechanisms in *B. gravinae*, *D. erucoides*, and *H. incana* HIR3. The carbon-concentrating mechanism in *D. muralis* is assumed to be inherited from the C_3–C_4 parent *D. tenuifolia* during natural hybridization with the C_3 species *D. viminea* (Eschmann-Grupe *et al.*, 2003; Ueno *et al.*, 2006).

Interestingly, from the two taxa assigned as *H. incana*, only one (HIR3) showed C_3–C_4-like features such as a CCP below 40 ppm and the typical organelle arrangement in the BS cells. Comparison of chloroplast sequences from both accessions revealed that only HIR1 clustered together with other accessions of this species while HIR3 sequences clustered closer to *Sinapis pubescens* and *Brassica procumbens* (Guerreiro *et al.*, 2023). This suggests that HIR3 and *H. incana* belong to different species, and the former represents a new C_3–C_4 lineage in the *Brassiceae*.

Brassicaceae *display large variation in efficiency of the carbon conservation mechanism but no C_4-like shuttles*

Our survey of CO_2-concentrating mechanisms in the *Brassicaceae* confirmed that measurements of the CCP represent a valuable tool for the identification of C_3–C_4 intermediate plant accessions. In agreement with the large CCP screening study by Krenzer *et al.* (1975) and our own statistical analysis, taxa with CCPs between 10 ppm and 40 ppm were classified as C_3–C_4 intermediates. In this group, we observed gradual changes in the CO_2-concentrating capacity. Our study therefore supports models claiming that after establishment of the basic photorespiratory shuttle, multiple metabolic and anatomic adjustments can contribute to the efficiency of the pathway, resulting in additive small fitness gains (Heckmann *et al.*, 2013).

The lowest CCPs in the present investigation were measured in *D. tenuifolia* and the C_3–C_4 *Moricandia* species. Although various accessions of these species were used in different studies (Hylton *et al.*, 1988; Rawsthorne *et al.*, 1988a; Apel *et al.*, 1997; Ueno *et al.*, 2003, 2006; Schlüter *et al.*, 2017), low CCPs seem to be a ubiquitous trait of these respective species. Moreover, low CCPs in these species were supported by BS-specific localization of the GLDP protein (Hylton *et al.*, 1988; Rawsthorne *et al.*, 1988a; Ueno *et al.*, 2003). Further, especially in *D. tenuifolia*, CCP values were observed as very low and close to those

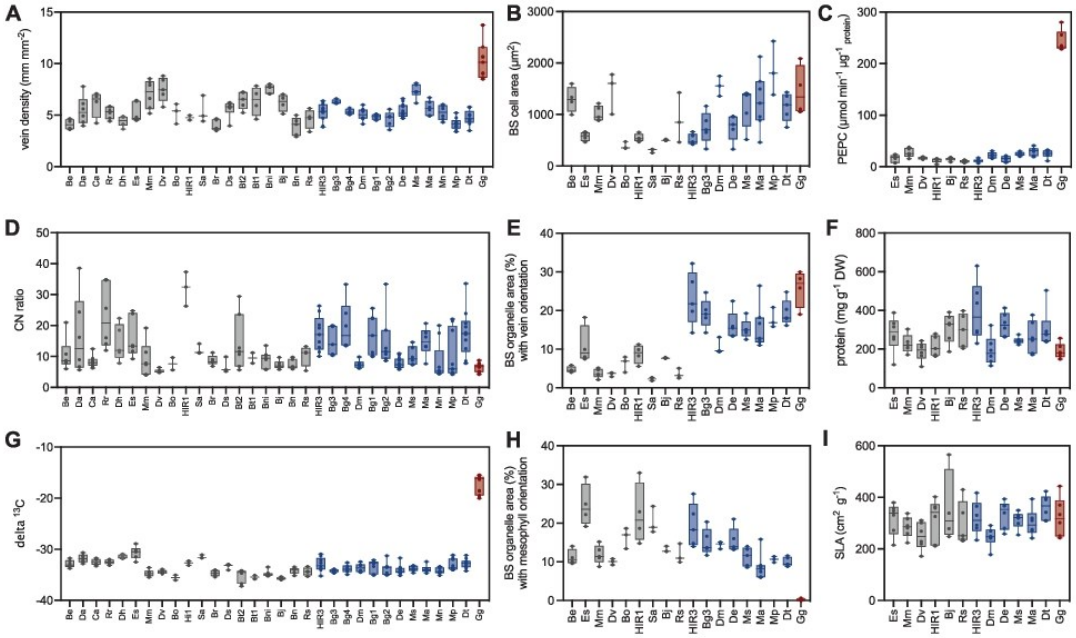

Fig. 8. Leaf structure- and composition-related parameters and PEPC activity. Mature leaves were used for determination of vein density (A), bundle sheath cell area in micrographs (B), PEPC activity (C), carbon to nitrogen ratio (D), area of bundle sheath organelles with vein orientation in micrographs (E), protein content (F), ^{13}C signature (G), area of bundle sheath organelles with orientation to intercellular space or mesophyll in micrographs (H), and specific leaf area (I). The tested taxa were sorted according to their CO_2 compensation point and coloured according to photosynthesis type as C_3 (grey), C_3–C_4 (blue), and C_4 (red). Taxa names have been abbreviated for legibility and are provided in Fig. 2 and the Materials and methods.

typical of C_4 species. However, the strict separation of the C_4 *G. gynandra* in all PCAs and especially the PEPC and ^{13}C measurements support previous observations claiming absence of C_4-like shuttles in the *Brassicaceae* (Holaday and Chollet, 1984; Hunt *et al.*, 1987; Sage *et al.*, 2011).

Reduction in CCP correlates negatively with organelle accumulation and arrangement in the BS

Despite the differences in efficiency of the photorespiratory shuttle, changes in organelle arrangement were observed in all taxa classified as C_3–C_4. For instance, all C_3–C_4 taxa possessed an enhanced BS area occupied by organelles in the centripetal position and a higher total organelle area per BS cell compared with C_3 species (Figs 8, 9). This underlines the importance of anatomical features for carbon-recapturing mechanisms. A strong correlation between reduction in CCP and increased organelle accumulation facing the vein in the BS was also previously observed in interspecific hybrids between *D. tenuifolia* (C_3–C_4) and *R. sativus* (C_3) (Ueno *et al.*, 2003). The BS structural features appeared to be genetically encoded and are inherited independently from the GLDP localization (Ueno *et al.*, 2003). Residual expression of *GLDP* was also observed in M cells of C_3–C_4 intermediate *Flaveria* species (Schulze *et al.*, 2013). This shows that structural modifications can underpin an effective CCP without complete suppression of GLDP in the M cells.

C_3–C_4 intermediates in our study contained several layers of M cells such that many do not directly border BS cells. So complete absence of GDC activity in the M cells would require transport of photorespiratory glycine through other M cell layers prior to entering the BS for metabolization. However, accumulation of mitochondria in the BS might create a glycine sink supporting glycine diffusion to the BS, and partial reduction of M GLDP expression would enforce the shuttle. TEM studies of centripetally localized organelles from C_3–C_4 *Brassicae* (Ueno *et al.*, 2006; Schlüter *et al.*, 2017), *Asteraceae* (McKown and Dengler, 2007), *Boraginaceae* (Muhaidat *et al.*, 2011), *Scrophulariaceae* (Khoshravesh *et al.*, 2012), *Arthropogoninae* (Khoshravesh *et al.*, 2016), and *Chenopodiaceae* (Yorimitsu *et al.*, 2019) have shown a close arrangement of mitochondria and chloroplasts. Thus, BS ultrastructure seems to play a major role in prevention of photorespiratory CO_2 and NH_3 loss and in improvement of leaf carbon and possibly also nitrogen economy.

In contrast to the C_4 species in our study, BS cells in C_3–C_4 *Brassicaceae* exhibited organelles facing the ICS and M cells (Fig. 8; Supplementary Fig. S6). This amount of ICS/M cell-facing organelles decreased in C_3–C_4 species with higher carbon-concentrating

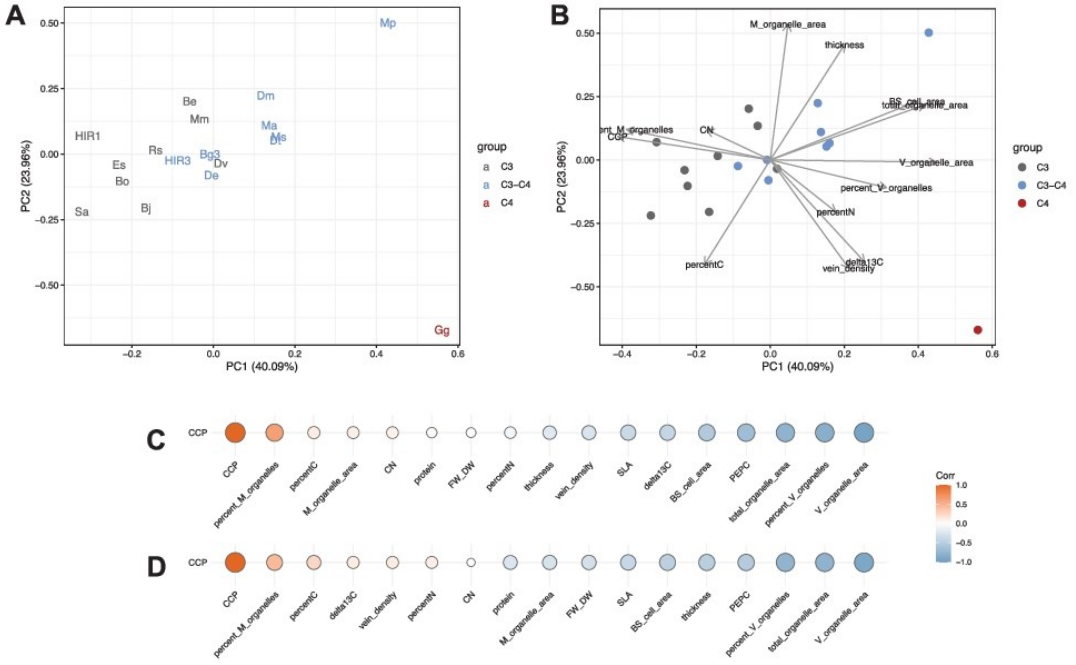

Fig. 9. Principal component analysis (PCA) and correlation of the CO_2 compensation point (CCP) with leaf structural and compositional components. Average values for the selected parameters measured by EA-IRMS analysis of leaf cross-sections by light microscopy. (A) Localization of the taxa in the PCA, (B) PCA including parameter loadings, (C) Pearson correlation coefficients demonstrated as heatmaps using all plant lines, (D) Pearson correlation coefficients demonstrated as heatmaps using only C_3 and C_3–C_4 lines. The tested taxa were coloured according to photosynthesis type as C_3 (grey), C_3–C_4 (blue), and C_4 (red). Taxa names have been abbreviated for legibility and are provided in Fig. 2 and the Materials and methods.

efficiency. Our results suggest that accumulation of centripetal organelles and reduction of peripheral organelles are not necessarily regulated by the same process. Additional structural features of C_4 species such as enlarged BS cell area and higher vein density did not differ between the tested C_3 and C_3–C_4 *Brassicaceae* taxa. Further, leaf thickness, SLA, and FW/DW ratios were also not different between the leaves of the C_3, C_3–C_4, and C_4 taxa (Fig. 8; Supplementary Fig. S8). Thus, despite leaf anatomy and BS architecture being important requirements for evolution of carbon-concentrating mechanisms (Christin *et al.*, 2013), modifications in leaf succulence parameters do not appear to be essential for an efficient photorespiratory carbon-concentrating pathway. Plasticity in some morphological parameters could also play a role in further evolution towards the C_4 leaf, and it could be speculated that limited genetic potential for the adjustment of vein density and mesophyll structure could be connected to the absence of C_4-like shuttles in the *Brassicaceae*.

Brassicaceae *C_3–C_4 metabolism had only a minor influence on leaf steady-state metabolite patterns*

Beside organelle arrangement in the BS, the shift of GDC activity to this tissue influences leaf biochemistry (Rawsthorne, 1992; Schlüter *et al.*, 2017). Relocation of the GLDP protein to the BS has been observed in all investigated C_3–C_4 classified species to date (Schlüter and Weber, 2016) and is therefore seen as the decisive step for the evolution of a photorespiratory carbon concentration shuttle. In the *Brassicaceae*, BS specificity of GLDP was shown for different *Moricandia* species (Rawsthorne *et al.*, 1988a), *Diplotaxis tenuifolia* (Ueno *et al.*, 2003), and *Brassica gravinae* (Ueno, 2011). Lack or reduction of GDC activity in the M causes accumulation of photorespiratory glycine and its transport along the concentration gradient to the BS. The GDC reaction converts two molecules of glycine into one molecule of serine, but also utilizes NADH and releases NH_3 alongside CO_2. This imbalance requires further metabolic readjustment between the two cell types. Nevertheless, beyond GLDP localization, not much is known about the cell-specific metabolism or the nature of additional metabolite shuttles in C_3–C_4 *Brassicaceae*.

If metabolite exchange between M and BS cells is realized by a concentration gradient, high concentrations of these transported metabolites would be expected in the leaves (Leegood and von Caemmerer, 1989). However, it should be noted that high metabolic flux and cell-specific metabolite accumulation might mask these gradients in total leaf extracts. Our metabolite analysis did not identify preferential metabolite shuttles

operating across all C_3–C_4 *Brassicaceae* species. Steady-state glycine concentrations were generally enhanced in the C_3–C_4 species compared with C_3 species, supporting the hypothesis that glycine is transported from the M cells to the BS for decarboxylation. High glycine was, however, also found in leaves of the C_3 *D. tenuisiliqua* and the C_4 species *G. gynandra*, indicating that glycine accumulation is not a distinct C_3–C_4 feature (Fig. 7A). Further uncertainty exists around the metabolites transported back from the BS to the M for rebalancing of carbon, nitrogen, and energy metabolism (Borghi *et al.*, 2022). Beside glycine, serine accumulation also exhibited a negative correlation with CCP values (Fig. 6D). This strongly supports the involvement of serine as a metabolite transported back from the BS to the M cells (Rawsthorne, 1992; Mallmann *et al.*, 2014), although variation in serine levels suggests that the contribution of serine transport could vary between the different taxa.

High variation between the individual taxa also existed for other shuttle metabolite candidates. Modelling approaches have previously predicted the involvement of glutamate, α-ketoglutarate, alanine, pyruvate, aspartate, and malate in shuttling processes for rebalancing of nitrogen metabolism between the M and BS (Mallmann *et al.*, 2014). Malate and aspartate could also be involved in rebalancing of reducing power between the two cell types (Johnson *et al.*, 2021). Contributions of glutamine/glutamate and asparagine/aspartate to intercellular shuttles were suggested for the C_3–C_4 species *Flaveria anomala* (Borghi *et al.*, 2022). Here, enhanced levels of these various metabolites could be observed in some, but not all, C_3–C_4 taxa (Fig. 7). For example, high concentrations of malate, aspartate, and glutamate were found in species displaying very low CCPs such as *M. arvensis* and *D. tenuifolia*. Interestingly, the C_3–C_4 *Moricandia* species which supposedly share a single C_3–C_4 evolutionary origin also showed strong variation in the metabolite pattern. A similar absence of main shuttle metabolites has also been described for C_3–C_4 *Flaveria* species (Borghi *et al.*, 2022). Our data generally support the hypothesis that multiple metabolites are transported between the M and BS (Schlüter *et al.*, 2017; Borghi *et al.*, 2022). The contribution of the different metabolites could differ in the individual taxa depending on genetic as well as environmental influences. Such a multitude of solutions indicates that metabolite and energy balancing does not represent a limiting step during evolution of carbon-concentrating pathways.

To date, enzyme localization studies have mostly focused on the GLDP protein, and much less is known about whether other reactions are shifted to the BS in C_3–C_4 species. In *M. arvensis*, other tested photorespiratory enzymes such as glycolate oxidase, serine hydroxymethyl transferase, and other subunits of the GDC complex were present in both cell types (Morgan *et al.*, 1993). Enzyme activities in *M. arvensis* in M- and BS-enriched fractions were also equally distributed for glyoxylate aminotransferases, glycolate oxidase, and hydroxypyruvate reductase (Rawsthorne *et al.*, 1988b), supporting the crucial role of GLDP for uneven distribution for glycine shuttle operation in *M. arvensis*. On the other hand, all GDC subunits were preferentially expressed in the BS in C_3–C_4 *Flaveria* and *Panicum* species (Morgan *et al.*, 1993). Shifting of additional photorespiratory steps could considerably influence the metabolite shuttles. In our study, some species, especially *D. encoides* and *D. tenuifolia*, showed high levels of glycolate and glycerate. Interestingly, intercellular transport of glycerate and glycolate was predicted in a constraint-based modelling approach for weak carbon-concentrating mechanisms on the evolutionary path to C_4 photosynthesis (Blätke and Bräutigam, 2019). Exchange of these metabolites between M and BS would reduce the need for intercellular nitrogen recycling (Borghi *et al.*, 2022). Part of the photorespiratory metabolites could also feed into additional pathways in the BS. It has been estimated that 1–5% of the photorespiratory glycine and ~30% of serine can be metabolized outside the photorespiratory cycle in processes such as protein biosynthesis (Busch *et al.*, 2018). The high organelle accumulation would increase the demand for protein synthesis in the C_3–C_4 BS. Furthermore, the BS is also responsible for loading of assimilation products into the phloem, and part of the carbon and nitrogen transported into the BS by the glycine shuttle could support metabolite export to the sink tissue of the plants.

C_3–C_4 photosynthesis is associated with reduced Ci and enhanced WUE especially under limiting CO_2

In the *Brassicaceae*, the presence of C_3–C_4 metabolism did not translate into improved photosynthetic assimilation under ambient environmental conditions (Figs 3, 4). For instance, across the *Brassicaceae* species analysed in the current study, assimilation rates appeared to be genotype specific rather than related to photosynthesis type under ambient CO_2. This lack of correlation between assimilation and photosynthesis type has also been previously described in the *Chenopodiaceae* (Yorimitsu *et al.*, 2019).

Interestingly, however, C_3–C_4 taxa in the *Brassicaceae* adjusted leaf *Ci* to lower levels compared with C_3 taxa in this clade. The difference between these photosynthesis types was marginal under ambient conditions, but became more pronounced under CO_2 conditions of ≤200 ppm (Fig. 3). The ability to assimilate CO_2 at lower *Ci* translated into higher WUE in the C_3–C_4 taxa compared with the C_3 species. This increase in WUE observed was underpinned by enhanced assimilation, as stomatal conductance was similar among the C_3 and C_3–C_4 taxa under all tested conditions (Fig. 4). It should be noted, however, that the differences observed for *Ci* and WUE between C_3 and C_3–C_4 taxa were small in comparison with the difference between all C_3 and C_3–C_4 taxa and C_4 *G. gynandra*, thus underlining the superiority of the C_4 pathway as a CO_2-concentrating mechanism compared with C_3–C_4 metabolism. Similar observations have been previously made in *Heliotropium* and *Flaveria*, in which C_3–C_4 species achieved WUE values between those of the C_3 and C_4 species. This was

also due to higher assimilation rather than modified conductance (Huxman and Monson, 2003; Vogan *et al.*, 2007). These results support an advantage of the C_3–C_4 pathway in high photorespiratory conditions which cause CO_2 restriction due to stomatal closure.

Evolution of the glycine shuttle often appears to be connected to an enlargement of the growth habitat (Lundgren and Christin, 2017). The C_3 species *M. moricandioides* for instance seems to be geographically restricted to the Iberian Peninsula, while the closely related C_3–C_4 species *M. arvensis* has spread into northwest Africa, Southern Europe, and other parts of the planet where it is mostly associated with cultivated areas and disturbed sites (Perfectti *et al.*, 2017). *Diplotaxis tenuifolia* also often grows as an invasive species occupying sunny, harsh, and arid areas (Nicoletti *et al.*, 2007) in which water, nutrient, and temperature conditions can change rapidly. It is therefore possible that C_3–C_4 species profit from high environmental plasticity of the trait.

Ecological studies which have investigated the adaptation of C_3–C_4 species to specific environmental conditions are unfortunately still rare (Oono *et al.*, 2022). In contrast to C_3 and C_4 species, the C_3–C_4 compensation points are strongly influenced by environmental conditions, especially light, temperature, and nitrogen (Brown and Morgan, 1980; Holaday and Chollet, 1983; Hunt *et al.*, 1987; Schuster and Monson, 1990; Gomez *et al.*, 2020; Oono *et al.*, 2022). In *Chenopodium album*, the CCP was lowest under high temperature and low nitrogen conditions, which was connected to accumulation of the GLDP protein preferentially in the BS (Oono *et al.*, 2022). *Moricandia arvensis* leaves also had lower CCPs and higher WUE under hotter and more arid summer conditions than in milder spring climates (Gomez *et al.*, 2020). Plasticity of photosynthetic traits under stress conditions was recently also reported for the C_2 species *Sasola divaricate* (Tefarikis *et al.*, 2022). Our results indicate that gradual and even facultative implementation of carbon shuttles between the M and BS are possible and should be considered in future experiments.

Knowledge about the distribution of species with glycine shuttle metabolism is generally still limited to studies among relatives of C_4 species. This is mainly due to the dependence on gas exchange equipment and time-consuming measurements. It is therefore assumed that the frequency of species with weaker carbon-concentrating mechanisms is greatly underestimated (Sage *et al.*, 2011; Lundgren, 2020). Identification of C_3–C_4 features in a *H. incana* HIR3 and recently also in some *C. album* accessions (Yorimitsu *et al.*, 2019) supports this hypothesis. As such, faster methods for identification of C_3–C_4 intermediates could help to close this knowledge gap. Here, our correlation analysis showed that measurements of assimilation at low CO_2 are sufficient for detection of C_3–C_4 phenotype and would save considerable time as opposed to having to calculate CCP by measuring assimilation across a range of CO_2 concentrations (Fig. 5D). For example, a very strong positive correlation in the present results was found to exist between CCP and assimilation rate at 50 ppm CO_2, which is close to the CCP of C_3 species. High and significant negative correlation to CCP also existed for WUE under CO_2 conditions of ≤200 ppm. As in our experiments assimilation generally correlated positively with photosynthesis efficiency F_v'/F_m', fluorescence combined with stomatal conductance measurements could possibly also be used in a fast initial screening experiments for identification of C_3–C_4 intermediates in the future.

Conclusions

Our survey revealed that photorespiratory shuttles evolved up to five times in the *Brassiceae* tribe in different genetic backgrounds. Measurements of the CCP indicated considerable variation in the pathway in the different tested taxa. Reduction in CCP was generally associated with organelle arrangement in the BS. Thus, elucidation of regulatory mechanisms underlying organelle multiplication and arrangement in the BS appear to be crucial for engineering an efficient glycine shuttle pathway into the leaf.

Although CCPs as low as 12 ppm were observed in *D. tenuifolia*, there was no evidence for the operation of C_4-like shuttles in the tested taxa, supporting its classification as a distinct pathway. All C_3–C_4 classified taxa belong to the *Brassiceae* tribe which appears to have lost one GLDP gene copy, suggesting that this event facilitated evolution of the glycine shuttle (Schlüter *et al.*, 2017). Additional loss-of-function mutations or insertion of a transposable element are thought to be involved in loss or reduction of GDC activity in the M cells (Rawsthorne, 1992; Sage *et al.*, 2012; Adwy *et al.*, 2015; Triesch *et al.*, 2023, Preprint). In *D. muralis*, transfer of weak carbon-concentrating mechanisms seem to have been inherited during hybridization from a C_3–C_4 parent (Ueno *et al.*, 2003). The contribution of hybridization to distribution of carbon-concentrating pathways has been discussed for several plant groups including *Sasola* and *Flaveria* (Kadereit *et al.*, 2017; Tefarikis *et al.*, 2022; Morales-Briones and Kadereit, 2023, Preprint). In some grasses, lateral gene transfer has been shown to support the rapid and successful establishment of the C_4 pathway (Dunning *et al.*, 2019b). Such scenarios would nevertheless require donor species that are able to successfully transfer essential features into the receiving genetic background.

Our results reveal that photorespiratory carbon-concentrating mechanisms in the *Brassiceae* show large variation in their biochemical and physiological features. C_3–C_4 *Brassiceae* species are often associated with fast changing temperature, water, and nutrient conditions. Metabolic plasticity could also be advantageous in crop species challenged by dynamic climatic variability. *Brassica napus* or *B. oleaceae* are closely related to the described C_3–C_4 species and would be prime targets for transfer of this trait. Recent progress in sequencing the genomes of these species and related species in the *Brassicaceae* (Guerreiro *et al.*, 2023) can help to identify the molecular mechanisms behind BS-specific C_3–C_4 architecture and biochemistry.

Supplementary data

The following supplementary data are available at *JXB* online.

Table S1. Origin of the seed material.

Table S2. Physiological data measured with the Li-COR 6800 (average per accession, standard deviation, HSD group).

Table S3. Metabolite data from GC-MS measurements (average per accession, standard deviation, HSD group).

Table S4. Data from light microscopy, EA-IRMS, and PEPC enzyme assay (average per accession, standard deviation, HSD group).

Fig. S1. Box and whisker plot for gas exchange and fluorescence data summarized per photosynthesis type.

Fig. S2. Heatmap of Pearson correlation matrix for physiological gas exchange and fluorescence data.

Fig. S3. Heatmap of Pearson correlation matrix for CO_2 compensation points and metabolite data from GC-MS analysis.

Fig. S4. Box and whisker plot for metabolite data summarized per photosynthesis type.

Fig. S5. Vein pattern in de-stained leaves.

Fig. S6. Micrographs of bundle sheath cross-sections.

Fig. S7. Box and whisker plot for structural and leaf composition data summarized per photosynthesis type.

Fig. S8. Heatmap of Pearson correlation matrix for CO_2 compensation point, structural parameters from light microscopy analysis, leaf composition data from EA-IRMS analysis. and PEPC enzyme activity.

Acknowledgements

The authors would like to thank the gardener team for looking after our plants in the HHU greenhouses. The technical assistance of Katrin Weber, Maria Graf, Elisabeth Klemp, and Dominik Brilhaus in the metabolite analysis is also greatly appreciated. We are grateful to Stephanie Krey for assistance with DNA extraction, and Perlina Lim for help during the setup of the microscopy experiments. Sebastian Triesch is thanked for his helpful comments on the manuscript. We also thank two unknown reviewers for valuable inputs and improvement of the manuscript.

Author contributions

APMW, BS, and US: initiation and planning of the experiments; US, JWB, MM, and CK: contributing data on plant physiology, biochemistry, and structure; PW: measurement and analysis of the plant metabolites; RG and BS: providing sequence data and the phylogenetic tree; US: writing the manuscript; JWB, MM, RG, BS, PW, and APMW: editing the manuscript. All authors read and approved the final manuscript.

Conflict of interest

No conflict of interests declared.

Funding

This work was funded by the ERA-CAPS project 'C4BREED' under Project ID WE 2231/20-1. It was further supported by the Cluster of Excellence for Plant Sciences (CEPLAS) under Germany's Excellence Strategy EXC-2048/1 under project ID 390686111, the H2020 EU project 'Gain4crops', and the CRC TRR 341 'Plant Ecological Genetics' grant by the German Research Foundation (DFG).

Data availability

The raw SRA data used for the phylogenetic tree are deposited in NCBI under BioProject PRJNA905373. Physiological, biochemical and anatomical data per taxon used in this study are available in Supplementary Tables S2–S4.

References

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

Apel P, Horstmann C, Pfeffer M. 1997. The *Moricandia* syndrome in species of the Brassicaceae. Photosynthetica **33**, 205–215.

Arias T, Pires JC. 2012. A fully resolved chloroplast phylogeny of the brassica crops and wild relatives (Brassicaceae: Brassiceae): novel clades and potential taxonomic implications. Taxon **61**, 980–988.

Atkinson RRL, Mockford EJ, Bennett C, Christin P, Spriggs EL, Freckleton RP, Thompson K, Rees M, Osborne CP. 2016. C4 photosynthesis boosts growth by altering physiology, allocation and size. Nature Plants **2**, 1–5.

Aubry S, Smith-Unna RD, Boursnell CM, Kopriva S, Hibberd JM. 2014. Transcript residency on ribosomes reveals a key role for the *Arabidopsis thaliana* bundle sheath in sulfur and glucosinolate metabolism. The Plant Journal **78**, 659–673.

Bellasio C. 2017. A generalized stoichiometric model of C_3, C_2, C_2+C_4, and C_4 photosynthetic metabolism. Journal of Experimental Botany **68**, 269–282.

Bellasio C, Farquhar GD. 2019. A leaf-level biochemical model simulating the introduction of C2 and C4 photosynthesis in C3 rice: gains, losses and metabolite fluxes. New Phytologist **223**, 150–166.

Blätke M-A, Bräutigam A. 2019. Evolution of C4 photosynthesis predicted by constraint-based modelling. eLife **8**, e49305.

Borghi GL, Arrivault S, Günther M, *et al*. 2022. Metabolic profiles in C_3, C_3–C_4 intermediate, C_4-like, and C_4 species in the genus *Flaveria*. Journal of Experimental Botany **73**, 1581–1601.

Brautigam A, Kajala K, Wullenweber J, *et al*. 2011. An mRNA blueprint for C4 photosynthesis derived from comparative transcriptomics of closely related C3 and C4 species. Plant Physiology **155**, 142–156.

Brown RH, Morgan JA. 1980. Photosynthesis of grass species differing in carbon dioxide fixation pathways.:VI. Differential effects of temperature and light intensity on photorespiration in C_3, C_4, and intermediate species. Plant Physiology **66**, 541–544.

Busch FA, Sage TL, Cousins AB, Sage RF. 2013. C_3 plants enhance rates of photosynthesis by reassimilating photorespired and respired CO_2. Plant, Cell & Environment **36**, 200–212.

Busch FA, Sage RF, Farquhar GD. 2018. Plants increase CO_2 uptake by assimilating nitrogen via the photorespiratory pathway. Nature Plants **4**, 46–54.

Campbell MS, Law MY, Holt C, *et al*. 2014. MAKER-P: a tool kit for the rapid creation, management, and quality control of plant genome annotations. Plant Physiology **164**, 513–524.

Christin P-A, Osborne CP, Chatelet DS, Columbus JT, Besnard G, Hodkinson TR, Garrison LM, Vorontsova MS, Edwards EJ. 2013.

Anatomical enablers and the evolution of C_4 photosynthesis in grasses. Proceedings of the National Academy of Sciences, USA **110**, 1381–1386.

Craine JM, Lee WG, Bond WJ, Williams RJ, Johnson LC. 2005. Environmetal constraints on a global relationship among leaf and root traits of grasses. Ecology **86**, 12–19.

Dunning LT, Moreno-villena JJ, Lundgren MR, *et al*. 2019a. Key changes in gene expression identified for different stages of C_4 evolution in *Alloteropsis semialata*. Journal of Experimental Botany **70**, 3255–3268.

Dunning LT, Olofsson JK, Parisod C, *et al*. 2019b. Lateral transfers of large DNA fragments spread functional genes among grasses. Proceedings of the National Academy of Sciences, USA **116**, 4416–4425.

Edwards GE, Ku MSB. 1987. Biochemistry of C_3–C_4 intermediates. In: Stumpf P, Conn E, eds. The biochemistry of plants. New York: Academic Press, 275–325.

Eilbeck K, Moore B, Holt C, Yandell M. 2009. Quantitative measures for the management and comparison of annotated genomes. BMC Bioinformatics **10**, 67.

Ellinghaus D, Kurtz S, Willhoeft U. 2008. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinformatics **9**, 18.

Emms DM, Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biology **16**, 157.

Emms DM, Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biology **20**, 1–14.

Eschmann-Grupe G, Hurka H, Neuffer B. 2003. Species relationships within *Diplotaxis* (Brassicaceae) and the phylogenetic origin of *D. muralis*. Plant Systematics **243**, 13–29.

Fiehn O, Kopka J, Dörmann P, Altmann T, Trethewey RN, Willmitzer L. 2000. Metabolite profiling for plant functional genomics. Nature Biotechnology **18**, 1157–1161.

Flügel F, Timm S, Arrivault S, Florian A, Stitt M, Fernie AR, Bauwe H. 2017. The photorespiratory metabolite 2-phosphoglycolate regulates photosynthesis and starch accumulation in Arabidopsis. The Plant Cell **29**, 2537–2551.

Gomez JM, Perfetti F, Armas C, Narbona E, Gonzalez-Megias A, Navarro L, DeSoto L, Torices R. 2020. Within-individual phenotypic plasticity in flowers fosters pollination niche shift. Nature Communications **11**, 4019.

Gowik U, Westhoff P. 2011. The path from C3 to C4 photosynthesis. Plant Physiology **155**, 56–63.

Gowik U, Bräutigam A, Weber KL, Weber APM, Westhoff P. 2011. Evolution of C_4 photosynthesis in the genus *Flaveria*: how many and which genes does it take to make C_4? The Plant Cell **23**, 2087–2105.

Gu Y. 2012. Multicomponent reactions in unconventional solvents: state of the art. Green Chemistry **14**, 2091–2128.

Guerreiro R, Bonthala VS, Schlüter U, Triesch S, Weber APM, Stich B. 2023. A genomic panel for studying C3–C4 intermediate photosynthesis in the Brassiceae tribe. Plant, Cell & Environment https://doi.org/10.1111/pce.14662.

Han Y, Wessler SR. 2010. MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. Nucleic Acids Research **38**, e199.

Heckmann D, Schulze S, Denton A, Gowik U, Westhoff P, Weber APM, Lercher MJ. 2013. Predicting C_4 photosynthesis evolution: modular, individually adaptive steps on a Mount Fuji fitness landscape. Cell **153**, 1579–1588.

Holaday AS, Chollet R. 1983. Photosynthetic/photorespiratory carbon metabolism in the C3–C4 intermediate species, *Moricandia arvensis* and *Panicum milioides*. Plant Physiology **73**, 740–745.

Holaday AS, Chollet R. 1984. Photosynthetic/photorespiratory characteristics of C3–C4 intermediate species. Photosynthesis Reseach **5**, 307–323.

Hunt S, Smith AM, Woolhouse HW. 1987. Evidence for a light-dependent system for reassimilation of photorespiratory CO_2, which does not include a C_4 cycle, in the C_3–C_4 intermediate species *Moricandia arvensis*. Planta **171**, 227–234.

Huxman TE, Monson RK. 2003. Stomatal responses of C_3, C_3–C_4 and C_4 *Flaveria* species to light and intercellular CO_2 concentration: implications for the evolution of stomatal behaviour. Plant, Cell & Environment **26**, 313–322.

Hylton CM, Rawsthorne S, Smith AM, Jones DA, Woolhouse HW. 1988. Glycine decarboxylase is confined to the bundle-sheath cells of leaves of C_3–C_4 intermediate species. Planta **175**, 452–459.

Janacek SH, Trenkamp S, Palmer B, *et al*. 2009. Photosynthesis in cells around veins of the C_3 plant *Arabidopsis thaliana* is important for both the shikimate pathway and leaf senescence as well as contributing to plant fitness. The Plant Journal **59**, 329–343.

Johnson JE, Field CB, Berry JA. 2021. The limiting factors and regulatory processes that control the environmental responses of C_3, C_3–C_4 intermediate, and C_4 photosynthesis. Oecologia **197**, 841–866.

Kadereit G, Bohley K, Lauterbach M, Tefarikis DT, Kadereit JW. 2017. C_3–C_4 intermediates may be of hybrid origin—a reminder. New Phytologist **215**, 70–76.

Khoshravesh R, Akhani H, Sage TL, Nordenstam B, Sage RF. 2012. Phylogeny and photosynthetic pathway distribution in *Anticharis* Endl. (Scrophulariaceae). Journal of Experimental Botany **63**, 5645–5658.

Khoshravesh R, Stinson CR, Stata M, Busch FA, Sage RF, Ludwig M, Sage TL. 2016. C_3–C_4 intermediacy in grasses: organelle enrichment and distribution, glycine decarboxylase expression, and the rise of C_2 photosynthesis. Journal of Experimental Botany **67**, 3065–3078.

Kinsman EA, Pyke KA. 1998. Bundle sheath cells and cell-specific plastid development in Arabidopsis leaves. Development **125**, 1815–1822.

Koch MA, Lemmel C. 2019. Zahora, a new monotypic genus from tribe Brassiceae (Brassicaceae) endemic to the Moroccan Sahara. PhytoKeys **135**, 119–131.

Krenzer EG, Moss DN, Crookston RK. 1975. Carbon dioxide compensation points of flowering plants. Plant Physiology **56**, 194–206.

Leegood RC. 2008. Roles of the bundle sheath cells in leaves of C3 plants. Journal of Experimental Botany **59**, 1663–1673.

Leegood RC, von Caemmerer S. 1989. Some relationships between contents of photosynthetic intermediates and the rate of photosynthetic carbon assimilation in leaves of *Zea mays* L. Richard. Planta **178**, 258–266.

Long SP. 1999. Environmental responses. In: Sage RF, Monson RK, eds. C4 plant biology. San Diego: Academic Press, 215–249.

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

Lundgren MR, Christin P. 2017. Despite phylogenetic effects, C_3–C_4 lineages bridge the ecological gap to C_4 photosynthesis. Journal of Experimental Botany **68**, 241–254.

Lundgren MR, Christin PA, Escobar EG, Ripley BS, Besnard G, Long CM, Hattersley PW, Ellis RP, Leegood RC, Osborne CP. 2016. Evolutionary implications of C_3–C_4 intermediates in the grass *Alloteropsis semialata*. Plant, Cell & Environment **39**, 1874–1885.

Lundgren MR, Osborne CP, Christin P. 2014. Deconstructing Kranz anatomy to understand C_4 evolution. Journal of Experimental Botany **65**, 3357–3369.

Mallmann J, Heckmann D, Bräutigam A, Lercher MJ, Weber APM, Westhoff P, Gowik U. 2014. The role of photorespiration during the evolution of C4 photosynthesis in the genus Flaveria. eLife **3**, e02478.

McKown AD, Dengler NG. 2007. Key innovations in the evolution of Kranz anatomy and C4 vein pattern in Flaveria (Asteraceae). American Journal of Botany **94**, 382–399.

Monson RK, Edwards GE, Ku MS. 1984. C_3–C_4 intermediate photosynthesis in plants. Biosciences **34**, 563–574.

Moore BD, Ku MSB, Edwards GE. 1989. Expression of C4-like photosynthesis in several species of Flaveria. Plant, Cell & Environment **12**, 541–549.

Morales-Briones DF, Kadereit G. 2023. Exploring the possible role of hybridization in the evolution of photosynthetic pathways in Flaveria (Asteraceae), the prime model of C_4 photosynthesis evolution. BioRxiv doi:10.1101/2022.01.31.478436. [Preprint].

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

Muhaidat R, Sage TL, Frohlich MW, Dengler NG, Sage RF. 2011. Characterization of C_3–C_4 intermediate species in the genus *Heliotropium* L. (Boraginaceae): anatomy, ultrastructure and enzyme activity. Plant, Cell & Environment **34**, 1723–1736.

Nicoletti R, Raimo F, Miccio G. 2007. *Diplotaxis tenuifolia*: biology, production and properties. The European Journal of Plant Science and Technology **1**, 36–43.

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

Perfectti F, Gómez JM, González-Megías A, Abdelaziz M, Lorite J. 2017. Molecular phylogeny and evolutionary history of *Moricandia* DC (Brassicaceae). PeerJ **5**, e3964.

Rawsthorne S. 1992. C_3–C_4 intermediate photosynthesis: linking physiology to gene expression. The Plant Journal **2**, 267–274.

Rawsthorne S, Hylton CM, Smith AM, Woolhouse HW. 1988a. Photorespiratory metabolism and immunogold localization of photorespiratory enzymes in leaves of C_3 and C_3–C_4 intermediate species of *Moricandia*. Planta **173**, 298–308.

Rawsthorne S, Hylton CM, Smith AM, Woolhouse HW. 1988b. Distribution of photorespiratory enzymes between bundle-sheath and mesophyll cells in leaves of the C_3–C_4 intermediate species *Moricandia arvensis* (L.) DC. Planta **176**, 527–532.

Roach MJ, Schmidt S, Borneman AR. 2018. Purge haplotigs: synteny reduction for third-gen diploid genome assemblies. BMC Bioinformatics **19**, 460.

Sage RF, Christin PA, Edwards EJ. 2011. The C_4 plant lineages of planet Earth. Journal of Experimental Botany **62**, 3155–3169.

Sage RF, Khoshravesh R, Sage TL. 2014. From proto-Kranz to C_4 Kranz: building the bridge to C_4 photosynthesis. Journal of Experimental Botany **65**, 3341–3356.

Sage RF, Sage TL, Kocacinar F. 2012. Photorespiration and the evolution of C_4 photosynthesis. Annual Review of Plant Biology **63**, 19–47.

Sage TL, Sage RF. 2009. The functional anatomy of rice leaves: implications for refixation of photorespiratory CO_2 and efforts to engineer C_4 photosynthesis into rice. Plant and Cell Physiology **50**, 756–772.

Sayyari E, Mirarab S. 2016. Fast coalescent-based computation of local branch support from quartet frequencies. Molecular Biology and Evolution **33**, 1654–1668.

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

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

Schlüter U, Weber AP. 2020. Regulation and evolution of C_4 photosynthesis. Annual Review of Plant Biology **71**, 183–215.

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

Schuster WS, Monson RK. 1990. An examination of the advantages of C_3–C_4 intermediate photosynthesis in warm environments. Plant, Cell & Environment **13**, 903–912.

Sharkey TD. 1988. Estimating the rate of photorespiration in leaves. Physiologia Plantarum **73**, 147–152.

Shim SH, Lee SK, Lee DW, Brilhaus D, Wu G, Ko S, Lee CH, Weber APM, Jeon JS. 2020. Loss of function of rice plastidic glycolate/glycerate translocator 1 impairs photorespiration and plant growth. Frontiers in Plant Science **10**, 1726.

Smit A, Hubley R, Green P. 2015. RepeatMasker Open-3.0. http://www.repeatmasker.org.

Taniguchi YY, Gowik U, Kinoshita Y, Kishizaki R, Ono N, Yokota A, Westhoff P, Munekage YN. 2021. Dynamic changes of genome sizes and gradual gain of cell-specific distribution of C_4 enzymes during C_4 evolution in genus Flaveria. Plant Genome **14**, 1–14.

Tefarikis DT, Morales-Briones DF, Yang Y, Edwards G, Kadereit G. 2022. On the hybrid origin of the C_2 *Salsola divaricata* agg. (Amaranthaceae) from C_3 and C_4 parental parents. New Phytologist **234**, 1876–1890.

Triesch S, Denton AK, Buchmannn JP, Reichel-Deland V, Guerreiro R, Schlüter U, Weber APM. 2023. Transposable elements contribute to the establishment of the glycine shuttle in Brassicaceae species. BioRxiv doi:10.1101/2022.12.06.519256. [Preprint].

Ueno O. 2011. Structural and biochemical characterization of the C_3–C_4 intermediate *Brassica gravinae* and relatives, with particular reference to cellular distribution of Rubisco. Journal of Experimental Botany **62**, 5347–5355.

Ueno O, Bang SW, Wada Y, Kondo A, Ishihara K, Kaneko Y, Matsuzawa Y. 2003. Structural and biochemical dissection of photorespiration in hybrids differing in genome constitution between *Diplotaxis tenuifolia* (C_3–C_4) and radish (C_3). Plant Physiology **132**, 1550–1559.

Ueno O, Wada Y, Wakai M, Bang SW. 2006. Evidence from photosynthetic characteristics for the hybrid origin of *Diplotaxis muralis* from a C_3–C_4 intermediate and a C_3 species. Plant Biology **8**, 253–259.

Vogan PJ, Frohlich MW, Sage RF. 2007. The functional significance of C_3–C_4 intermediate traits in *Heliotropium* L. (Boraginaceae): gas exchange perspectives. Plant, Cell &Environment **30**, 1337–1345.

Williams BP, Johnston IG, Covshoff S, Hibberd JM. 2013. Phenotypic landscape inference reveals multiple evolutionary paths to C_4 photosynthesis. eLife **2**, e00961.

Yorimitsu Y, Kadosono A, Hatakeyama Y, Yabiku T, Ueno O. 2019. Transition from C_3 to proto-Kranz to C_3–C_4 intermediate type in the genus *Chenopodium* (Chenopodiaceae). Journal of Plant Research **132**, 839–855.

Zhang C, Scornavacca C, Molloy EK, Mirarab S. 2020. ASTRAL-pro: quartet-based species-tree inference despite paralogy. Molecular Biology and Evolution **37**, 3292–3307.

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

Here I helped Prof. Stich supervise the initial bioinformatical work of Dr. Triesch, providing him with the data and teaching him how to analyze it. The initial discovery of the correlation of M-box distance to the *GLDP* gene and CCP happened as a result of this supervision. We also started the investigation of binding sites in upstream areas of genes. Continued exchange of ideas remained. As a co-author, I reviewed the manuscript but made minimal text contributions.

plant biology

Plant Biology ISSN 1435-8603

RESEARCH ARTICLE

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

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

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

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

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

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

Editor
H.-H. Kunz

Received: 8 September 2023;
Accepted: 15 November 2023

doi:10.1111/plb.13601

ABSTRACT

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

INTRODUCTION

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

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

Changes in the spatial and temporal patterns of gene expression are crucial for the evolution of C_3-C_4 intermediate photosynthesis (Hibberd & Covshoff 2010; Reeves *et al.* 2017). Previously, it has been shown that the BSC-specific decarboxylation of glycine is caused by the differential

Plant Biology **26** (2024) 270–281 © 2024 The Authors. *Plant Biology* published by John Wiley & Sons Ltd on behalf of German Society for Plant Sciences, Royal Botanical Society of the Netherlands.

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

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

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

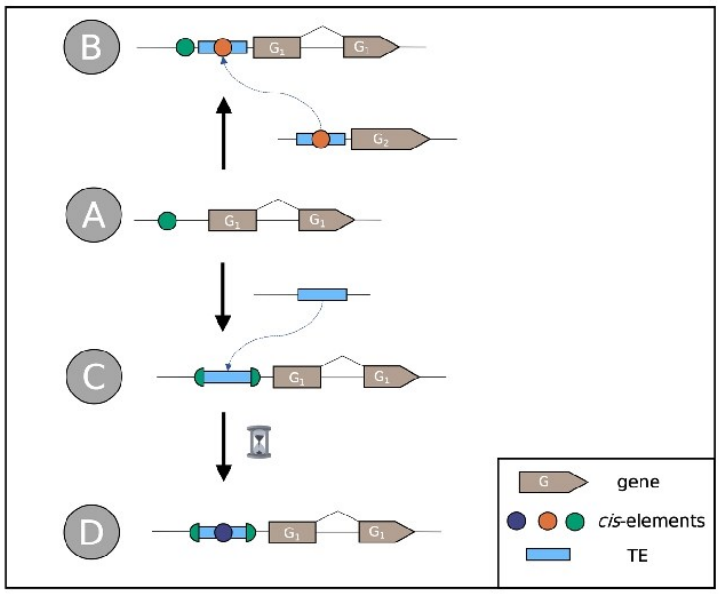

Fig. 1. Schematic illustration of gene regulation rewiring by TE exaptation. A: The hypothetical gene G1 is controlled by a *cis*-regulatory element (CRE, green dot). B: Gene G2 is regulated by a different CRE (orange dot) located within a TE (blue box). Upon transposition of the TE to the upstream region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

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

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

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

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

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

MATERIAL AND METHODS

Genomes and carbon compensation points

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

Gene annotation

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

Annotation of transposable elements

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

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

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)

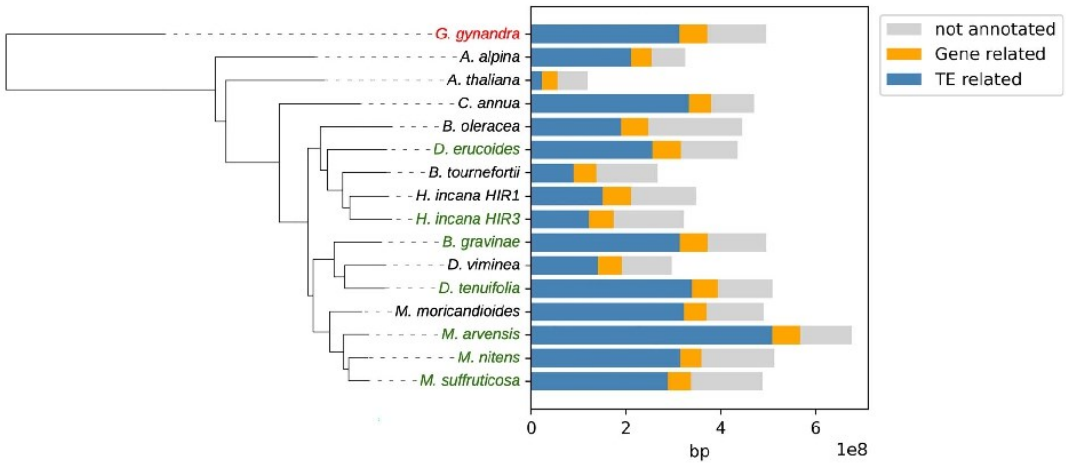

Fig. 2. Phylogeny and genomic composition of 15 selected Brassicaceae species and the Cleomaceae outgroup. C_3-C_4 intermediate species are highlighted in green, the C_4 outgroup *Gynandropsis gynandra* is highlighted in red. TE-related nucleotides are defined as spanning intact and fragmented transposon.

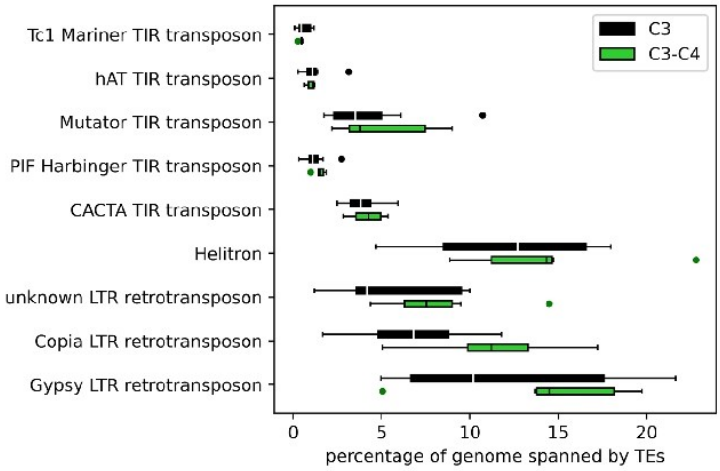

Fig. 3. Boxplot indicating the percentage of the genome comprised by each class of intact and fragmented TEs in eight C_3 and six C_3-C_4 intermediate species. The y-axis shows the TE classes, the x-axis indicates the fraction of the genome made up by the respective TE class. Black boxes depict C_3 species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

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

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

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

RESULTS

The TE landscape of C_3 and C_3-C_4 Brassicaceae species

To screen for genomic features of potential relevance to the evolution of the C_3-C_4 photosynthesis trait, we conducted a pan-genomic association study of eight C_3 Brassicaceae species,

seven C_3-C_4 intermediate Brassicaceae species from five independent origins, and one C_4 Cleomaceae as an outgroup species for tree building. The five independent origins of C_3-C_4 intermediate photosynthesis can be found in the *Moricandia arvensis*, *M. nitens*, and *M. suffruticosa* monophylum, as well as in *Diplotaxis erucoides*, *D. tenuifolia*, *Brassica gravinae*, and *Hirschfeldia incana* HIR3 (Fig. 2) (Schlüter & Weber 2016; Schlüter *et al.* 2022; Guerreiro *et al.* 2023).

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

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

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

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

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

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

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

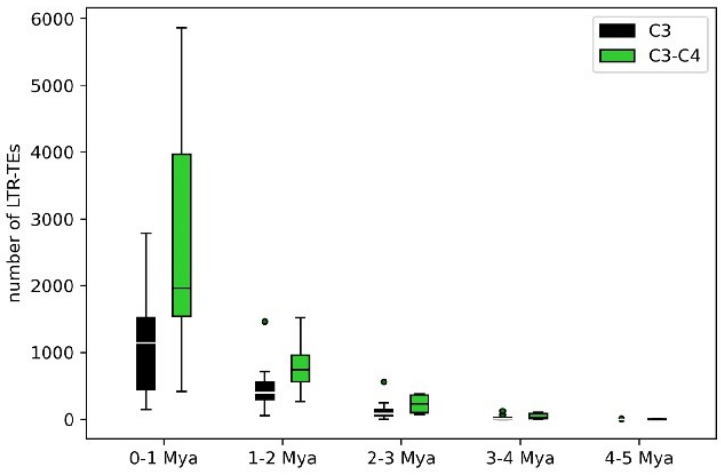

Fig. 4. Boxplot of LTR-TE insertion times for eight C_3 and six C_3-C_4 intermediate species. The x-axis shows the insertion time in bins of 1 million years before today (Mya). The y-axis depicts the number of identified LTR-TEs calculated to be inserted within this time frame. Calculation was performed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

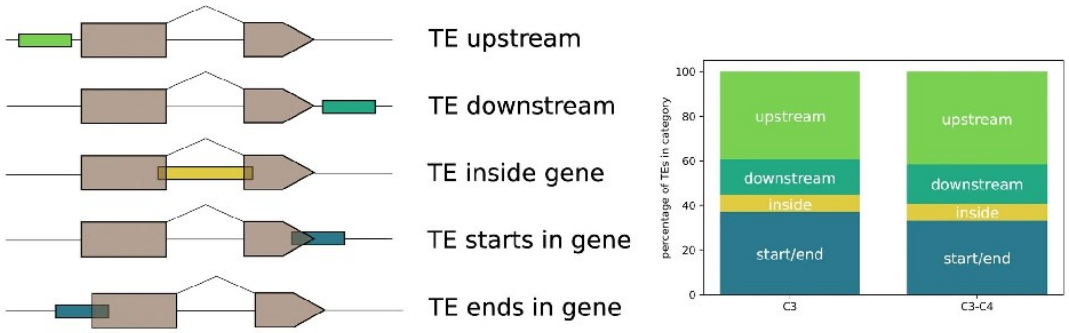

Fig. 5. Left panel: Different contexts of TEs co-occurring with genes. Right panel: Bar charts indicating the fractions of TE co-occurring with genes within five contexts: starting or ending in a gene (start/end), residing within a gene (inside) or residing within a 3000 bp window upstream or downstream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

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

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

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

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

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

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

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

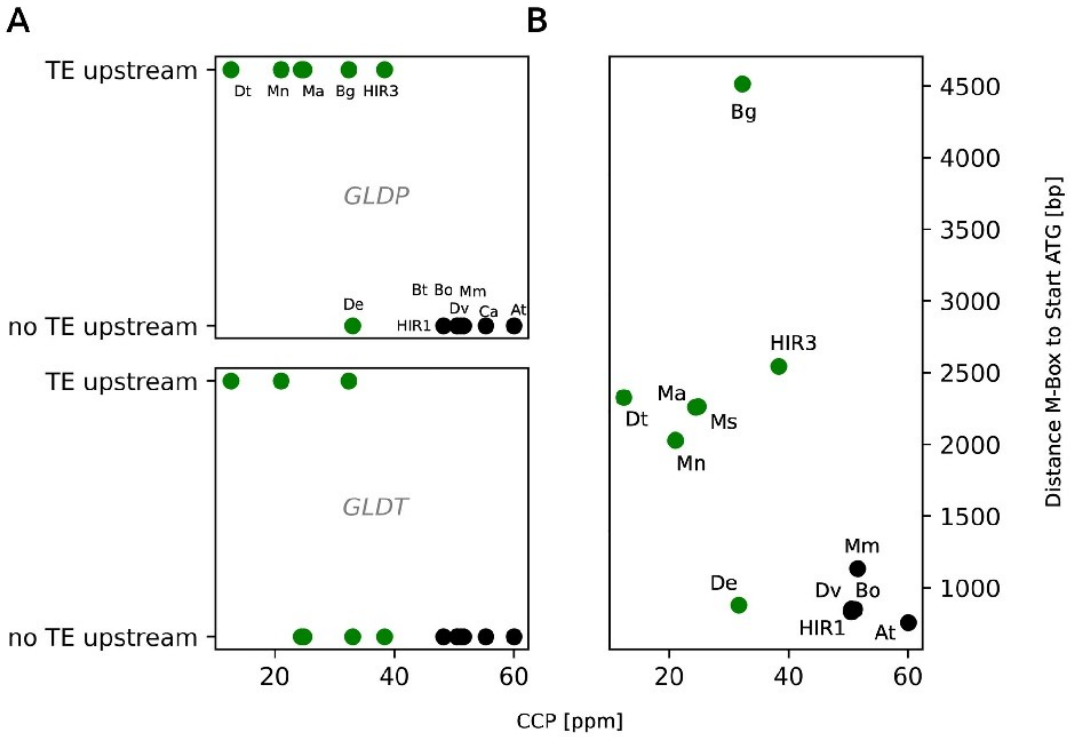

Fig. 6. A: Scatter plot for two photorespiratory genes with significant co-associated upstream TEs. The y-axis indicates the presence of an upstream TE (yes/no), the x-axis shows the carbon compensation point. Abbreviations: *GLDP*/*GLDT*: P/T-protein of the GLYCINE DECARBOXYLASE COMPLEX B: Scatter plot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

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

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

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

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

DISCUSSION

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

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

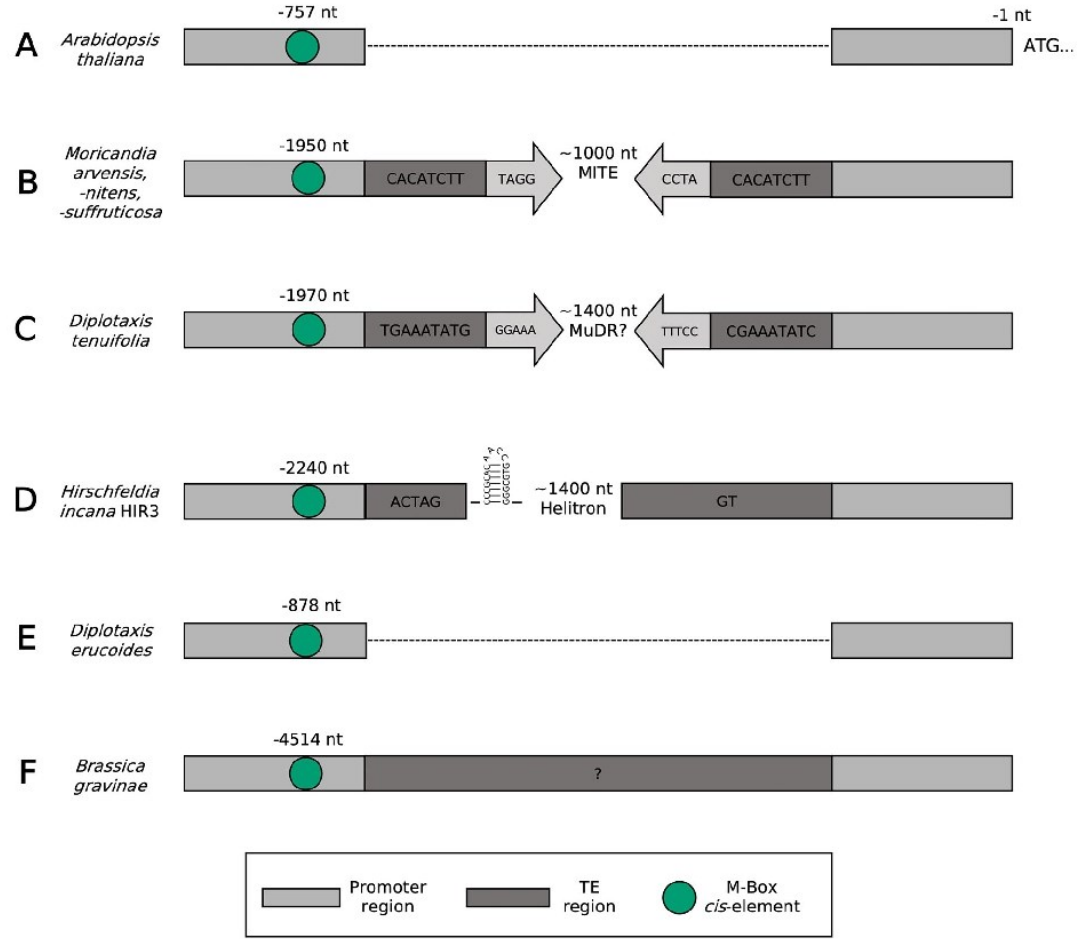

Fig. 7. Schematic representation of the *GLDP1* promoter region. "ATG…" depicts the start site of the *GLDP1* gene. Dark grey boxes represent characteristic TE sites such as target site duplications or the *Helitron* insertion sites. Grey arrows depict terminal inverted repeat motifs. The M-Box motif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

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

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

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

results and genome size estimations have to be handled with care.

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

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

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

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

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

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

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

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

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

Adwy *et al.* (2019) explain the establishment of the glycine shuttle in *Moricandia* by the loss of the M-Box in C_3-C_4 intermediate *Moricandia* species. However, in contrast to this, we

found the M-Box sequence in all our analysed *GLDP1* promoter variants, although this motif was shifted by over 1000 bp further upstream by the insertion of three independent TEs in the promoters in three independent evolutionary origins of C_3-C_4 intermediate photosynthesis, and by an insertion of unknown provenance in a fourth independent origin. This shift may have led to the M-Box being overlooked in previous studies.

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

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

ACKNOWLEDGEMENTS

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

AUTHOR CONTRIBUTIONS

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

SUPPORTING INFORMATION

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

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

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

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

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

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

Appendix S2. Supporting Information.

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

REFERENCES

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

new tools. *Nucleic Acids Research*, **40**(D1), D1202–D1210.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Discussion

The work of this thesis is intended on establishing a groundwork for investigation of C_3-C_4 photosynthesis in the Brassiceae tribe of the Brassicaceae family. The set of species included in this study has been systematically characterized for C_3-C_4 traits, including Carbon Compensation Point (CCP), and other physiological and anatomical traits, revealing variation inside the family and between closely related species. The published set of genomes allows a robust inference of phylogenetic relatedness, as well as the analysis of variation of genes and their regulatory regions across several degrees of C_3-C_4 phenotype. The published articles have attracted attention of reviews (Smith et al., 2023) and insight articles (Walsh et al., 2023; Julkowska et al., 2024), highlighting the novelty of the genome assemblies and the relevance of the wild C_3-C_4 species for the introduction of desirable traits into commercial plants.

Bioinformatical approaches to study C_3-C_4 in Brassicaceae

Genome assembly and annotation

A total of 19 genome assemblies created within this thesis are new to science, exhibiting varying levels of contiguity. While commonly used as a proxy for assembly quality, contiguity measures alone are not reliable for quality assessment in a genome assembly (Simão et al., 2014; Thrash et al., 2020). In that regard, the assemblies largely capture the reference gene space of the Eudicot clade (Guerreiro et al., 2023, Fig. 1). The resulting total assembly sizes, representative but not fully equal to the real species' genome sizes, have no correlation with the C_3-C_4 phenotype (Triesch et al., 2023, Fig.2). While the assemblies have contiguity considered sufficient for gene annotation, comparable to other Brassicaceae species, there is potential for contiguity and haplotype-phasing improvement through future sequencing. This improvement could be achieved with long-read-based sequencing as exemplified here with 7 species (Guerreiro et al., 2023, Suppl. Table 1), or with proximity-ligation sequencing, as exemplified in the annex paper for Potato (Freire et al., 2021). This was in the meantime done for both HIR1 and HIR3 accessions of *Hirschfeldia incana*, improving assembly through long read sequencing (El Hasnaoui et al., 2024).

The *de novo* gene annotation strategy yielded an average of 46,546 high-quality gene models

across taxa, comparable to publicly available gene annotations for related species. Importantly, despite imperfect levels of assembly contiguit*y,* a high proportion of annotated genes are accompanied by long upstream sequences (>30Kb) (Guerreiro et al., 2023, Suppl. Fig. 5), enabling deeper analysis cis-regulatory motifs and transcription factors, ultimately offering valuable insights into gene regulation mechanisms. However, the noted exceptions in lower contiguity assemblies of *B. gravinae*, *D. harra*, and *M. sinaica* highlight challenges in genomic data consistency, underscoring the need for improved sequencing techniques or assembly methods to fully leverage genomic information. Lower quality assemblies may lead to missing or unconnected parts of the genome, whereby genes or their upstream areas might be falsely accounted (Thrash et al., 2020). This can disrupt various kinds of downstream comparative analysis as is visible in Guerreiro et al., 2023, where the less contiguous assemblies stand out in orthology grouping (Suppl. Fig. 7) and synteny (Fig. 5; Suppl. Fig. 7).

The importance of selfing, the creation of individuals from a self-fertilized parent individual, is evident for quality genome assemblies in plants. The most contiguous assemblies in Guerreiro et al. (2023) were the ones that used more selfed specimens, with the best *D. tenuisiliqua* being an F2 generation (Guerreiro et al., 2023, Suppl. Table 1). Selfing reduces heterozygosity between homologous chromosome sets, allowing an easy assembly of their sequences for a single haploid assembly. This is especially important for polyploid species and was an essential part of the annex paper on tetraploid potato (Freire et al., 2021). Of course, intra-species diversity may be lost with selfing, but it can easily be represented as called single-nucleotide polymorphisms (SNPs) or called structural variants (SVs), by mapping reads of heterozygous individuals to the haploid assembly, as illustrated by the second annex paper (Schmidt et al., 2024). In any case, C_3-C_4 traits are thought to be under positive selection pressure (Edwards et al., 2014) and therefore likely fixed in the species where they exist (Booker et al., 2017). If traits are not fixed in all haplotypes of a species, or if nucleotide variation exists in key genes, the strength and frequency of positive selection pressure for the trait may be lower (Booker et al., 2017), or it might be an example of balancing selection (Hedrick, 2007).

For further refined insights, the level of analysis must move to the gene level. To assist that, a new generation of annotation methods are set to further streamline comparative studies, making the detection of genes, repeats and other genomic features easier to run, more insightful and efficient (Stiehler et al., 2020). Indeed, the genome annotation was one of the most time consuming parts of this thesis (Guerreiro et al., 2023) and would nowadays be considerably

faster to perform with deep learning pipelines like Helixer (Holst et al., 2023). Finally, a holistic analysis of gene variation should start by grouping genes based on homology, and by estimating phylogenetic relations between them and between the species at hand.

Orthology and phylogenetic relations

Orthology grouping of genes across genome assemblies is a foundational step in comparative genomics, identifying homologous genes with probable common ancestry (Emms & Kelly, 2019). Orthologous genes likely retain similar functions across different organisms and comparing them offers significant insights into gene evolution (Gabaldón & Koonin, 2013). Chiefly, gene orthology helps transferring functional annotations from model organisms to less-studied species, accelerating gene discovery and functional validation efforts (Klasberg et al., 2016). During the work of this thesis, I developed and shared an add-on pipeline (https://github.com/davidemms/OrthoFinder/issues/451) for streamlining the interpretation of orthology groups, based on annotations from a reference like Araport11.

The 42 928 orthogroups identified in Guerreiro et al. (2023) represent structural and functional relations of genes across this species set, providing a level of abstraction necessary for hypothesis inference and testing. The grouping of genes should be robust, a recent comparison on Brassicaceae concluded that different orthology inference algorithms predict similar orthogroups (Liao et al., 2024). Identifying candidate genes involved in C_3-C_4 photosynthesis pathways can be done not only via orthogroup presence/absence across species (Nagy et al., 2014), but also via copy number variation (McCarroll et al., 2007). Moreover, even with well-known genes of interest, the detailed study of sequence diversity in orthologs or in their upstream areas may reveal variation of biological relevance, not in the least related to gene regulation. For instance, Cao et al. (2016) used 40 gene orthologs between maize and rice to identify BS specific motifs that differ between C_3 and C_4 photosynthesis. Likewise, an illustration is given in this thesis: the genes from the newly produced assemblies were grouped via orthology (Guerreiro et al., 2023). After annotating the presence of TEs in the upstream areas of those genes, a correlation could be identified between TE presence and C_3-C_4 photosynthesis for 222 orthogroups (Triesch et al., 2023). One of those orthogoups represented the *GLDP* genes, which under scrutiny led to conclusions of biological relevance, discussed in the "Genetic regulation mechanisms" section of this thesis' discussion.

By grouping related genes, orthology enables creating phylogenetic trees of gene relatedness

(Emms & Kelly, 2019). Concatenating the trees of thousands of genes results in high fidelity species phylogenies when using advanced coalescent models like Maximum Likelihood and Bayesian methods (Degnan & Rosenberg, 2009; Trifinopoulos et al., 2016). Often phylogenetic gaps inhibit the exploration of evolutionary hypotheses (Julkowska et al., 2024). The phylogeny established in Guerreiro et al. (2023) crucially provides the structure to analyze evolutionary pathways of C_3-C_4 evolution within the Brassiceae tribe, suggesting that the transition towards C_3-C_4 intermediacy has occurred independently five times. This revelation, identifying an additional origin than previously recognized, paints a more complex picture of C_3-C_4 photosynthesis in the tribe. It does not yet confirm or disprove the hypothesis that the C_2 carbon pump evolutionary endpoint on its own (Lundgren, 2020; Walsh et al., 2023), though it anecdotally reinforces that theory. It also offers further datapoints that contradict Kadereit et al. (2017)'s hypothesis that C_3-C_4 is merely the result of hybridization with C_4 species, as there are no known C_4 species in this family (Sage et al., 2011), and even if one were to have existed in the past and gone extinct, the probability of hybridization with 5 species seems lower.

The current paradigm shift intends to move from single model species into model clades (Mabry et al., 2023). In that new paradigm, this thesis's characterization of the Brassicaceae clade not only refines our understanding of plant evolutionary biology but also holds implications for leveraging genetic diversity in crop improvement. While convergent innovations in all C_3-C_4 species are obvious candidates for missing pieces in an overall strategy towards C_4 genetic engineering, non-convergent innovations may be relevant towards understanding distinct evolutionary pathways. Perhaps several roads towards C_4 exist (Schlüter & Weber, 2016), and some lineages would rather benefit from genetic mechanisms extracted from closely related C_3-C_4 species instead than from more distant ones. Put in simple words: Commercial species like *Brassica rapa*, *B. napus*, *B. oleracea* should more easily benefit from the introduction of genetic features taken from C_3-C_4 *D. erucoides* or HIR3 *H. incana* than from C_4 Maize (*Zea mays*), or even from *D. tenuifolia,* which belongs to another clade inside Brassiceae (Guerreiro et al., 2023, Fig. 4).

Anatomy and physiology of C_3-C_4 photosynthesis in Brassicaceae

The physiological and anatomical insights from Schlüter et al. (2023) provide a robust framework to underline the nuances of C_3–C_4 intermediate photosynthesis in Brassicaceae. The

clinal variation of CCP values (Schlüter et al., 2023, Fig. 1) confirms that not all C_3–C_4 species are the same, and they probably have varying degrees of derivations from the primitive C_3 state. Despite variation, evidence points to no Brassicaceae species as having more advanced C_4-like characteristics, as shown by the low abundance of phosphoenolpyruvate carboxylase (PEPC) and typical C_4 metabolites α-ketoglutarate, pyruvate and α-alanin (Schlüter et al., 2023, Fig. 7G-I and Fig. 8C). Additionally, all C_3-C_4 plants here are more similar to C_3 than C_4 in vein density, BS cross-section area, as well as in having significant portions of BS cells with organelles oriented towards the mesophyll (Schlüter et al., 2023, Fig. 8A,B,H). Taken together with the known BS cell specificity of *GLDP* in many of these species (Rawsthorne et al., 1988a; Rylott et al., 1998; Ueno et al., 2003; Ueno, 2011; Schlüter et al., 2016), it seems like all these Brassicacaeae lineages have developed a C_2 carbon shuttle but did not advance much on the ladder of supposed inevitable evolution towards C_4 (Gowik & Westhof, 2011; Sage et al., 2014; Edwards, 2014).

Common characteristics of C_3-C_4

The one defining feature common across all C_3-C_4 plants in this study was enhanced organelle accumulation in BS cells on the side adjacent to the central vein (Schlüter et al., 2023, Fig. 8E). This is called centripetal accumulation and is here corroborated as a basal C_3-C_4 trait, as previously suggested (Gowik & Westhoff, 2011; Sage *et al*., 2014). Centripetal organelle accumulation is particularly significant because mitochondria are where the photorespiratory step of glycine decarboxylation occurs, catalyzed by GDC, which is limited to the BS cells in C_3-C_4 (Rawsthorne et al., 1988a, b; Rylott et al., 1998; Schulze et al., 2016). Furthermore, chloroplasts are where RuBisCO is located, which captures CO_2 after it is released by glycine cleavage (Sage et al., 2014; Khoshravesh et al., 2016). Having a high concentration of both these organelles in the centripetal pole of the BS cells probably minimizes loss of either glycine or CO_2 back to the mesophyll through dispersion, increasing the efficiency of the carbon capture mechanism (Khoshravesh et al., 2016; Schlüter et al., 2023). This pattern of organelle accumulation had previously been reported for interspecific hybrids of C_3 and C_3-C_4, being estimated to be genetically encoded (Ueno et al., 2003). However, Schlüter et al. (2023) suggests that centripetal accumulation of organelles might be separately regulated from their peripheral reduction (Fig. 8H). Identifying the convergent genomic features that enable this organelle distribution pattern should be now possible with this panel of species (Guerreiro et al., 2023). Genes thought to underlie chloroplast localization, *PHOT2* and *CHUP1*, have

already been observed to have differential spatial expression in *Moricandia arvensis*, based on promoter-GUS assays (Lin et al., 2022). A next step is extending that analysis to all species in this panel. Finally, though centripetal accumulation is increasingly seen as a solid feature of C_3-C_4, it may not be essential in all plant families, as it does not seem to affect CCP values in C_3-C_4 Amaranthaceae (Oono et al., 2022). Other quantified anatomical features in Schlüter et al., (2023) don't seem to correlate with CCP values (Fig. 8), but they also remain to be explored in relation to genetic background of this panel of species.

Advantages and evolution of C_3-C_4

Schlüter et al. (2023) reports carbon assimilation rate and water use efficiency of C_3-C_4 Brassicaceae to be at the same level as in C_3 in contemporary atmospheric conditions but shows that those statistics would be significantly better than C_3 in pre-industrial carbon concentration in the atmosphere (Fig. 3 and 4). This is in accordance with literature, suggesting a stronger selective pressure in past atmospheric conditions to evolve a C_2 carbon shuttle (Sage et al., 2014). Though it seems here like C_3-C_4 has little water use advantage over C_3 under the current atmosphere (Fig. 4), it is important to note that our experiment was performed at stable abiotic conditions, whereas it is known that C_3-C_4 metabolism can widely vary at different soil nitrogen, temperature, or light conditions (Brown & Morgan, 1980; Schuster & Monson; 1990; Oono et al., 2022). In fact, studies with Brassicaceae have already reported lower CCPs and higher water use efficiency in C_3-C_4 *Moricandia* under warm and dry conditions than in mild conditions (Gomez et al., 2020), and even increased vein density when growing under drought and heat stress (Zhu et al., 2022). Moving forward, a focus on metabolic plasticity is a must when studying C_3-C_4.

In C_3-C_4 and especially in C_4 species, coordination between mesophyll and BS is necessary to maintain functional stoichiometry in both cell types (Sage & Mckown, 2006). On the contrary, the undifferentiated photosynthetic cells of C_3 species have more autonomy to adjust to shifting abiotic conditions (Sage & Mckown, 2006). This means that C_3 photosynthesis is more flexible in its ecology, akin to the concept of generalist niche (Futuyma & Moreno, 1988), whereby C_4 photosynthesis with its high fitness in stable hot environments would fit the concept of specialist ecological niche (Sage & Mckown, 2006). C_3-C_4 photosynthesis seems to hit a particular balance between avoiding the photorespiratory problem in warm dry environments and dealing with shifting abiotic conditions. However, since the photorespiratory problem does not

significantly occur in northern latitudes, the ecological window where C_3-C_4 photosynthesis is potentially fitter than C_3 or C_4 photosynthesis is relatively small. Despite this, C_3-C_4 Brassicaceae have large geographical distributions, along less hot and dry environments than typical for C_4, with richer soils and abiotic variation (Lundgren & Christin, 2017). Indeed, many C_3-C_4 Brassicaceae are invasive across the world (Gorai et al., 2009; Dahlin et al., 2012; Abu-Dieyeh et al., 2013; Winkler et al., 2019), which typically happens with generalist species that have ecological plasticity to adapt well to different environments (Gioria et al., 2023). Particularly, a richer soil is important for C_3-C_4 species due to a lower efficiency of nitrogen use than C_4 species (Ghannoum *et al.*, 2009, 2011). But environments where chilling can happen drastically reduce the advantages of C_4 (Pignon et al., 2019). Additionally, low light intensity or light variability via shading from other plants has been reported to undercut C_4 efficiency (Sage & McKown, 2006; Pignon et al., 2017; Bellasio & Farquhar, 2019). The idea that the C_2 shuttle represents an evolutionary endpoint would be reinforced by a strategic trade-off, where metabolic plasticity is prioritized over peak photosynthetic efficiency at ideal conditions (Lundgren, 2020).

Diverse metabolic backgrounds

Schlüter et al. (2023) also explores the role of metabolic background in C_3–C_4 biochemistry. While previous research highlighted the transportation of various metabolites between BS and mesophyll in association with this photosynthesis type (Heckmann et al., 2013; Mallmann et al., 2014; Schlüter et al., 2017; Pinheiro et al., 2023), the work of this thesis only confirmed glycine and serine as universal C_3–C_4 metabolites in Brassicaceae (Schlüter et al., 2023, Fig. 6D). High glycine concentrations are evidence of the established C_2 shuttle, which in turn produces serine and ammonia when decarboxylated in the BS cells (Schulze et al., 2016). Interestingly, high serine levels seem to be a feature of only C_3–C_4 species, not being present in C_4 *G. gynandra* (Schlüter et al., 2023). While the retrotransportation of serine to mesophyll cells is known for nitrogen stoichiometry (Rawsthorne, 1992), it alone cannot prevent nitrogen imbalance (Mallmann et al., 2014). In the panel of analyzed Brassicaceae, the extent of serine transport and the concentration of alternative metabolites for nitrogen rebalancing varies per species (Schlüter et al., 2023, Fig. 7). For example, both *D. erucoides* and *D. tenuifolia* favor glycerate, complemented by malate in first and glutamate in the latter (Schlüter et al., 2023, Fig. 7). *M. arvensis* utilizes a combination of glutamate and aspartate (Schlüter et al., 2023, Fig. 7). Of note, previous research had pointed *M. arvensis* as not having significant levels of aspartate

(Schlüter et al., 2017), an inconsistency which reflects the metabolome's variability and reactivity to environmental factors (Sampaio et al., 2016), as well as the limitation of using whole leaf samples instead of single cell datasets. This conclusion can also be taken when looking at the disparate metabolite variation even inside *Moricandia* (Schlüter et al., 2023, Fig. 7), which should have a common origin of the phenotype and thus similar metabolite patterns. However, the same lack of a consistent single other metabolite for ammonia refixation and shuttling has been reported with C_3-C_4 *Flaveria* (Borghi et al., 2022). The forming consensus is that multiple metabolites contribute for nitrogen rebalancing between BS and mesophyll, so the metabolite patterns don't diverge evidently from C_3 (Leegood & von Caemmerer 1994; Borghi et al., 2022). Nonetheless, metabolites such as pyruvate and α-alanine, predicted by Mallmann et al. (2014) to form a C_4-like shuttle were not highly present in any of surveyed C_3-C_4 species (Schlüter et al., 2023, Fig. 7), reinforcing the idea that evolution did not go further than the C_2 shuttle in this family, multiple times. Walsh et al. (2023) argues against the nitrogen imbalance hypothesis, proposing advantages of C_2 in fluctuating environments due to stress resilience through a secondary open flux carbon cycle and photorespiration pathways. Because most of photorespiratory cycle in C_2 still happens openly in the mesophyll, its' metabolites can easily become substrates for other processes (Keys, 1999), conferring higher flexibility to deal with shifting conditions (Timm et al., 2012; Walker et al et al., 2000). Furthermore, the variation inside this family, especially the quite distinct *D. erucoides*, highlights the possible existence of subgroups inside C_3-C_4 (Schlüter et al., 2023, Fig. 7) which could potentially be specific adaptations to different environmental conditions.

Having primary metabolite candidates for C_3-C_4 species identified for each species, the next steps should focus on experimental validation to confirm their functional roles. Techniques such as isotopic labeling can be used to track the flux of these metabolites in the tissue (Gevaert et al., 2008), revealing their direct impact on photosynthesis and carbon transport. A deeper understanding of the shuttle mechanisms can also be reached through studying the enzymes responsible for metabolite production and interconversion. For instance, literature modeling predicts that several aminotransferases are highly expressed in C_3–C_4 leaf tissue to recycle ammonia released during carbon transport to bundle sheath cells (Mallmann et al., 2014). Research on C_4 plants has even shown that high aminotransferases expression happens on the BS cells and not mesophyll (Doring et al., 2016; Schlüter et al., 2018), but enzyme localization research on C_3–C_4 has mostly been limited to GLC protein complex (Schlüter & Weber, 2020; Oono et al., 2022). Recently, a study identified more transcript expression for aspartate

aminotransferase (*AspAT*), as well as dicarboxylate transporter 2 (*DiT2*) and phosphoenolpyruvate/phosphate translocator (*PPT*) in *M. suffruticosa* than in C_3 *Brassica napus* (Zhu et al., 2022). Similar procedures applied to the whole panel of species published here could be very enlightening. Further research steps would greatly benefit from the adoption of new technologies like single cell sequencing or immunohistochemistry microscopy (Nawy, 2014; Rao et al., 2021), or at the very least use RNA sequencing of partitions of the leaf, as done by Doring et al., 2016.

Genetic regulation mechanisms

Most C_3-C_4 research focuses on transcriptomics/metabolomics and differential expression of genes between species or individuals (Schlüter et al., 2017; Lauterbach et al., 2017; Pinheiro et al., 2023) and not between cell types. However, little has been done to pinpoint the exact genetic control mechanisms through which differential expression is achieved between cells (Schlüter & Weber, 2016). Moreover, RNA counts are not always directly translatable to fitness gains or even to increased presence of their corresponding protein (Evans et al., 2015). Examples of gene regulation differing between cell types exist for other characteristics, many of which result promoters and other regulatory elements in the genome (Abdolreza et al., 2014; Kim & Sung, 2017; Mhiri et al., 2022), but little genomic work has been applied to C_3-C_4 specifically. Scientific literature is increasingly aware of the role of transposable elements (TEs) as dynamic components of plant genomes, contributing to gene regulation, genome structure, and evolution (Feschotte, 2008). They can influence gene expression by inserting themselves near or within genes, creating new regulatory elements, and altering chromatin structure (Mhiri et al., 2022; Hassan et al., 2024). Research has already associated TEs to the evolution of photosynthesis pathways. Looking at 40 ortholog genes between C_3 rice and C_4 maize, Cao et al. (2016) identified over 1000 differentially distributed promoter motifs, more than 60% being associated with TEs. These motifs likely originated from non-photosynthetic genes and were transposed to C_4 genes, linking gene regulatory networks and underscoring the evolutionary impact of TEs (Cao et al., 2016). The work of this thesis suggests TEs as underlying a primary step in C_3-C_4 evolution, namely the institution of the glycine carbon (C_2) shuttle between two differentiated cell types in the leaf: The mesophyll and the bundle sheath (BS) cells. This shuttle is known to be controlled by differential expression of the GLDP protein between the two cell types (Monson et al., 1984; Rawsthorne et al., 1988a; Schulze et al., 2016). Adwy et al. (2015, 2019) proposed the absence of an M-Box element as the reason for differential expression in C_3-C_4

intermediate *Moricandia*. However, our quality genome assemblies allowed for the identification of that M-box in those same species, upstream of the GLDP1 gene, merely shifted further upstream than expected in the C_3-C_4 species (Triesch et al., 2023). Thus, all studied C_3-C_4 species, except for *D. erucoides,* have an M-box promoter which exists at a higher distance to the *GLDP* gene than the C_3 species (Triesch et al., 2023, Fig. 6A and 7). Our counter proposal is then that the distance between the M-Box and *GLDP* gene is what causes the gene to be differentially expressed, the exact mechanism to be uncovered but it potentially has to do with chromatin structure (Mhiri et al., 2022; Hassan et al., 2024). Other regulation mechanisms such as transcription factor activity, DNA methylation, histone modifications (Kim & Sung, 2017), and the presence of small RNAs (Zhan & Meyers, 2023) may not be excluded but seem less likely in this context. The exception of *D. erucoides*, as in the metabolite analysis, seems to again indicate that this species achieved low CCP values through an evolutionary process that is not totally convergent with the other species in this panel. If so, it should be an interesting path to scrutinize on its own, but it should be also weighted as an outlier, so not to lower the discovery of true positive signal for the other species in this panel.

Our hypothesis is that some TEs, inserted in specific locations, underlie spatially differential gene expression (Triesch et al., 2023, Fig. 6,7). The specific relevant TEs happen independently of global TE patterns, which do not correlate with the phenotype of interest (Triesch et al., 2023, Fig. 2,5). If anything, recent insertion times of LTR transposons may be connected to C_3-C_4 evolution, as they are significantly more pronounced in C_3-C_4 species than in C_3 species, especially in *Moricandia arvensis* (Triesch et al., 2023, Fig. 4, Table S3). In addition to the findings on the M-Box, Triesch et al. (2023) revealed significant correlations between TE insertions in cis-regulatory regions and C_3-C_4 intermediacy in 113 genes (Triesch et al., 2023, Fig. 6A, Table 1), which group functionally under photorespiration, SnRK1-kinase regulation, cell wall proteins and solute transport channels (Triesch et al., 2023, Table 2). The role of TEs in rewiring gene regulatory networks has been documented in other plant species, indicating a broader evolutionary mechanism (Feschotte, 2008). Understanding this, an avenue for genetic improvement of commercial crops starts being visible. Differential regulation of GLDP1 and other relevant genes through TE insertions could be a target for genetic engineering. By manipulating the positioning of TEs or associated cis-elements, it may be possible to replicate the gene regulatory networks observed in C_3-C_4 intermediates, thereby enhancing the efficiency of C_3 crops (Triesch et al., 2023).

Future avenues for C_3-C_4 research in *Brassicaceae*

Directions of research

The work related to this thesis contributed to advance the knowledge and resources related to C_3-C_4 photosynthesis in Brassicaceae. The availability of complete genome assemblies is basal for much of the research that is to come. It makes it easier to design arrays for probing genomic variability across populations, as well as to engage in genome editing for knock-out experiments (White et al., 2013) or for transformation with new genes (South et al., 2019). Relevant genetic variation between species can be identified through direct comparative genomics (Triesch et al., 2023) or by a genomic dissection of segregating hybrid populations with varying phenotypes (Schlüter & Weber, 2016), as illustrated by a companion paper of this thesis (Schmidt et al., 2024). This has been tried once with Brassicaceae crops: Zhang et al. (2004) produced tetraploid hybrids of C_3-C_4 *Moricandia* and C_3 *Brassica* species, further backcrossing them with the commercial C_3 parent and using genomic markers to distinguish whether offspring inherited the C_3 or C_3-C_4 version of *GLDP*. Developing and applying markers to the recombinant lines was one of the biggest challenges in the study (Zhang et al., 2004). Such limitations can finally be overcome, it is now easier and more effective to design high quantities of markers for many genes and regulatory regions.

The utility of these methods is further enhanced by the harnessing of *H. incana* HIR3, newly described as a C_3-C_4 intermediate species (Schlüter et al., 2023, Fig. 1). Being phylogenetically closer to many crop species than *Diplotaxis* or *Moricandia* — both already previously explored on hybridization (Ueno et al., 2003; Bang et al., 2007; Ueno et al., 2007) — *H. incana* offers a promising resource for future studies. This species has already garnered increased attention following the publication of its genome and C_3-C_4 status (Taylor et al., 2023; Hoang et al., 2024). Based on its' CCP values (Schlüter et al., 2023, Fig.1), the studied HIR3 accession seems to be on the earliest stages of C_3-C_4 evolution, which may be interesting to identify which are the minimal steps necessary for a functioning C_2 cycle. Unlike all other lineages where C_3-C_4 intermediacy was identified in this study, there are no studies on this species identifying differential expression of the *GLDP* gene in BS cells, although that is to be predicted based on the presence of the M-box promoter (Triesh et al., 2023). The HIR3 accession is quite unique, and the work of Guerreiro et al. (2023) hints at the possibility of it being a separate species from the regular HIR1 genotype of *H. incana*. This is backed by morphological differences

(Guerreiro et al., 2023, Fig. 6C) and by phylogenetic work based on chloroplast DNA, where the two genotypes are not monophyletic (Guerreiro et al., 2023, Fig. 6A-B). Guerreiro et al. (2023) places HIR3 closer to *Sinapsis pubescens* and *Brassica procubens* than to the HIR1 accession, which clusters together with the NIJ accession recently published by Garassino et al. (2022) (Guerreiro et al., 2023, Fig. 6B). Sequencing the genomes and characterizing photosynthetic properties like the CCP of *S. pubescens* and *B. procubens,* or of closely related species in literature like *Diplotaxis brachycarpa* (Warwick & Hall, 2009) would add value to this panel and to C_3-C_4 research in Brassicaceae.

Another interesting species in this thesis is *D. muralis*, thought to be a tetraploid hybrid of *D. tenuifolia* and *D. viminea,* for which the first genomic evidence is provided by this thesis (Guerreiro et al., 2023, Fig. 5). This hybrid species is especially interesting since one parent is C_3-C_4 and the other is C_3, and its phenotype is intermediate between the two types (Schlüter et al., 2023, Fig. 1). Allele specific expression analysis on inter-specific hybrids from C_3 and C_3-C_4 parents had already identified cis-regulatory differences in *Moricandia* in *GLDP1* and other genes (Lin et al., 2022). The same analysis may be extended to *D. muralis*, to other genes and even other species of this panel, provided that interspecific hybrids can be generated. If the creation of interspecific segregating material is possible, the genomes of the various progeny can be cheaply sequenced on low coverage, mapping and accessing what parts of DNA are inherited from the C_3 or C_3-C_4 ancestor. That information can be used against individual phenotypical measurements like CCP or anatomical measurements, allowing to pinpoint specific genomic regions that correlate with phenotype, with increased statistical power through replication (Schlüter & Weber, 2016). Those regions can then be further scrutinized for key genomic determinants of the C_3-C_4 phenotype, be it structural variation, single nucleotide polymorphisms, regulatory binding sites or factors associated with chromatin accessibility.

In the future, several research pathways will be necessary to fully understand the C_3-C_4 and C_4 systems. Besides the genes related to glycine shuttling and decarboxylation focused on this thesis, there are many other points where C_3-C_4 and C_4 photosynthesis are different to C_3: in RuBisCO and chloroplast expression and localization, stomatal behavior, vein density, leaf anatomy and nitrogen economy (Ku et al., 1983; Vogan & Sage, 2011; Griffiths et al., 2013; Stata et al., 2016; Oono et al., 2022). While much knowledge has been gathered in these areas, significant gaps remain, particularly in understanding how gene expression is regulated. Breakthroughs already exist — genetic mechanisms to circumscribe the localization of

RuBisCO and chloroplasts to BS cells have been identified (Bowman et al., 2013; van Rooijen, 2020) — but comparative genomics could help uncover additional regulatory elements and gene interactions that have not yet been explored, offering a broader framework to fully elucidate these complex systems.

Phylogenetic association mapping

This thesis also makes the case that comparing multiple species at the same time, considering their evolutionary relations, is more insightful than only pairwise comparisons. The published panel of species is ideal for phylogenetic association mapping, using mixed-linear models to statistically detect correlations between genomic features and phenotypic variability (Hiller et al., 2012; Kiefer et al., 2019). In this approach, a quantified phenotype of interest, such as the Carbon Compensation Point (CCP), is modelled as a function of both fixed and random effects. Specifically, for each genetic variant (i), the phenotype (y) is regressed against a matrix of genetic variation (X) as a fixed effect, and a species variance-covariance computed from phylogenetic relatedness (Z) as random effect, with an additional error term (ϵ):

$$y = \alpha + \left(X_i\beta_i + Zb_i\right) + \epsilon$$

Where α is a fixed intercept, and β_i and b_i are the coefficient vectors applied to the corresponding matrices on the i-th genetic variant. Additionally, population structure may be included as a fixed effect in the form of principal components of genetic variation, *PC1* and *PC2*, multiplied by corresponding coefficient vectors U_i and u_i :

$$y = \alpha + X_i\beta_i + PC_1U_i + PC_2u_i + Zb_i + \epsilon$$

Scripts for achieving this were developed in the course of this thesis and are published in a public github repository (https://github.com/ViriatoII/C4Evol/blob/master/PAM_brassicas.R). This model structure allows for the disentanglement of the genetic and evolutionary influences on the phenotype, providing a robust framework for identifying significant associations between genetic variation and traits while accounting for the underlying phylogenetic relationships. The matrix of genetic variation can be comprised of a number of things: Either a matrix of presence/absence or copy-number variation of genetic orthogroups, or a matrix of SNP/indel variation in the upstream areas of genes. In any case, due to high numbers of

variables in a genetic matrix, multiple testing correction strategies like Bonferroni (Dunn, 1961) or Benjamini-Hochberg (Benjamini & Hochberg, 1995) adjustment of the significance threshold are necessary. These strategies ensure that the associations identified are robust and reliable, reducing the likelihood of spurious results. Additional literature applying similar procedures can be found (Prudent et al., 2016).

The results of phylogenetic association mapping can guide downstream research into the molecular mechanisms underlying trait variability. Genetic variation identified as significantly associated with traits like CCP can act as an entry point for studying genes and their regulatory context. As illustration, the M-box element regulatory element of *GLDP1* should be flagged as correlated with CCP values (Triesch et al., 2023), and further research on the functional implications of that could ensue.

Downstream genomic research

The biggest question remaining in this thesis has to do with how differential regulation of the *GLDP1* gene is created via the larger distance between the gene and a promoter M-box element (Triesh et al., 2023). It seems that this is a case where chromatin structure plays a role in gene regulation. Studies of chromatin structure could become a final piece in solving the puzzle. Whilst the field is still in its infancy in plants, first studies have already described non-random changes to chromatin structure between plant species (Vergara et al., 2017; Ullah et al., 2018; Dong et al., 2020; Li et al., 2023). Differences in chromatin structure can have practical effects in gene expression, as chromatin folds into areas of high transcription accessibility or areas of intense interaction frequency, named topologically associated domains (TADs) (Beagan et al., 2020). As an illustration, Farmer et al. (2021) demonstrated distinct chromatin accessibility profiles between cell types in *Arabidopsis* roots to be associated with differential gene expression and cell-specific physiology. While chromatin structure was not scrutinized for Brassicaceae in this thesis, the synteny analysis of Guerreiro et al., 2023 (Fig. 5) already provides a framework of colinear blocks and synteny breaks, which often coincide with open-chromatin boundaries of TADs (Li et al., 2023). Ultimately however, data on the physical proximity of the folded DNA is required. A common method to obtain such data is proximity ligation, also called Hi-C sequencing (Lieberman-Aiden et al., 2009). Naturally, Hi-C sequencing would also greatly help with further scaffolding the genome assemblies into chromosome level contiguity, as demonstrated by Freire et al. (2021) in this dissertation. Deciphering gene regulation and chromatin dynamics may eventually also require the

exploration of epigenetic modifications to the genomes. Two key epigenetic mechanisms, DNA methylation and histone methylation, play pivotal roles in shaping the chromatin landscape and how genes are accessed for transcription. DNA methylation affects chromatin density by adding methyl groups to DNA, typically at CpG sites, which can suppress gene activity, being often studied through techniques such as bisulfite sequencing (Darst et al., 2010). Likewise, histone methylation patterns can affect how open or close the chromatin is, leading respectively to gene activation or repression, and are investigated through various methods, including chromatin immunoprecipitation followed by sequencing (ChIP-seq) (Mardis, 2007). The integration of these epigenetic studies would build a holistic understanding of the molecular mechanisms governing gene expression and chromatin organization. Indeed, a comprehensive integration of multi-omic data is necessary to fully uncover the complexity and adaptability of photosynthetic mechanisms within Brassicaceae, unlocking potential introduction of C_3-C_4 mechanisms in agricultural and ecological contexts. The full understanding of gene regulation and metabolic flexibility in this trait offers valuable insights for practical applications in sustainable agriculture, especially under the current challenges posed by environmental degradation. Any important findings in Brassicaeae may further be compared with the *Flaveria* genus, which is increasingly the subject of attention of C_3-C_4 photosynthesis research (Mallman et al., 2014; Borghi et al., 2022; Munekage & Taniguchi, 2022).

Crop engineering

Efforts to engineer efficiency gains in crops illustrate the difficulty of implementing a complex photosynthetic system. For instance, Wang et al. (2017) transformed rice plants using maize GLK genes with a ubiquitin promoter specific to BS cells, resulting in the accumulation of organelles in the BS cells, a proto-Kranz-like anatomy. However, this transformation did not yield significant improvements in photosynthetic performance compared to the wild type. Likewise, Jethva et al. (2024) managed to relocate carbonic anhydrase activity from chloroplasts to the cytosol, a typical C_4 trait, but also did not have any impact on plant growth, photosynthetic rate or CO_2 assimilation. Both these transformations implemented C_4 sub-components onto a C_3 plant without palpable agricultural benefit. There are proposals to test inducing the genetic mechanism that limits RuBisCO to BS cells onto Arabidopsis *thaliana* (Singh & Reeves, 2020), which should significantly decrease the amount of this enzyme

produced overall (Furbank & Taylor, 1995) and thus decrease the nitrogen requirements (Ghannoum & von Caemmerer, 2011; Bar-On & Milo, 2019). However, this too is a C_4-like characteristic that requires well defined Kranz-anatomy and a functional carbon shuttle towards the BS cells to have efficiency (Li et al., 2017), even decreasing plant resilience to chilling in relation to C_3 and C_2 types (Pignon et al., 2019). Due to the stepwise nature of C_4 evolution many of the more popular C_4-like traits compound on earlier steps (Heckmann et al., 2013). It is essential to build a foundational structure before any C_4 traits are to be tested, and that may pass by the engineering of a mature C_2 shuttle first, along with other basal C_3-C_4 characteristics. As far as literature search could tell, the only attempts to establish the C_2 glycine shuttle into a C_3 commercial species were the previously mentioned hybridization efforts of Zhang et al. (2004) on *Brassica* crops. It resulted in modest phenotypic gains, not reaching the CCP levels of the C_3-C_4 *Moricandia* parent. Hybridization with backcrossing is not only very labor-intensive, but it also introduces other unwanted DNA segments from the second parent (Gramazio et al., 2021), often even whole sets of chromosomes with polyploidization (Zhang et al., 2004). Modern tools like CRISPR/Cas9 make it possible to precisely incorporate only the desired C_3-C_4 genomic changes while preserving the identity of the crop species (Zhang et al., 2018). To my knowledge, all further attempts at transformation of a C_3 plant aim directly towards C_4 characteristics (Lundgren, 2020).

All these efforts would benefit from a solid groundwork of a successful transformation of *A. thaliana* into C_2, where consequences on plant resilience and parameters like CCP can be easily studied, before moving towards commercial crop transformation and measurement of yield improvements. Ongoing research in this area will doubtless continue to shed light on the genetic mechanisms underlying C_4 and C_3-C_4 intermediate photosynthesis, with potential implications for improving crop productivity and resistance in the face of global demand and climate change. Ultimately, simply introducing the C_2 glycine shuttle to commercial crops may not be sufficient to address agricultural challenges but is a critical intermediate step toward achieving the more ambitious goal of full C_4 photosynthesis.

Conclusion

The research here presented has significantly advanced our understanding of C_3-C_4 intermediate photosynthesis in the Brassicaceae family, shedding light on the genomic, physiological, and evolutionary mechanisms underlying this trait. By establishing a panel of Brassicaceae genomes with detailed genetic and phenotypic annotations, a robust foundation for future comparative studies was laid, allowing the exploration of the genetic basis of C_3-C_4 intermediacy.

The phylogenetic analysis conducted in this thesis has revealed multiple independent origins of C_3-C_4 photosynthesis inside the Brassicaceae family, suggesting a more complex evolutionary pathway than previously understood. Crucially, a thus far considered genotype of *Hirschfeldia incana*, described here for the first time as a C_3-C_4 intermediate, falls inside a sub-clade that was not known to possess C_3-C_4 species, in yet another apparent process of convergent evolution. Support for convergent evolution is found in increased Carbon assimilation rates and water use efficiency under past atmospheric conditions.

The analysis of metabolic pathways illustrated the complex diversity of metabolites involved in the C_3-C_4 transition, highlighting variability among species and potential metabolic plasticity. Furthermore, the investigation into genomic regulation mechanisms shed light on the influence of transposable elements on a crucial step of cell type differentiation of bundle sheath cells, through control of *GLDP* differential expression, challenging literature assumptions. Unlike previously thought, it is not the absence of a regulatory element that is correlated with differential expression, but the increased distance of that regulatory element to the gene, prompting hypothesis based on chromatin structure related regulation. This reframing may inform both future research decisions and targeted genetic engineering aimed at improving crop resilience and efficiency.

Looking ahead, the gained insights could be leveraged to enhance crop productivity and sustainability, particularly in the context of global food security and climate change. By reproducing the genetic characteristics and adaptive mechanisms found in C_3-C_4 intermediates, it may be possible to develop new crop varieties that boast higher photosynthetic efficiency. Introducing the simple C_2 glycine shuttle onto crops. Future research building on this work will undoubtedly continue to unravel the complexity of this trait and of plant evolution, paving the way for innovative strategies in crop improvement and sustainable agriculture.

Bibliography

Abogadallah, G.M. (2011). Differential regulation of photorespiratory gene expression by moderate and severe salt and drought stress in relation to oxidative stress. *Plant Science*, 180, 540–547.

Abdolreza, E. N., Fatemeh, E., Elyas, M. S., Ahmad, S. N., Maryam, G., Behzad, G., & Ali, M. M. (2014). Structures of Phosphoenolpyruvate carboxylase (PEPC) Gene Promoter from C_4 and C_3 *Flaveria* species Using Sequence Analysis by Bioinformatics Tools. *Annual Research & Review in Biology*, *4*(17), 2779-2794.

Abu-Dieyeh, M. H., Al-Abeid, H., & Qnais, E. Y. (2013). Does allelopathy increase the invasion of *Diplotaxis erucoides* in disturbed arid land?. *Allelopathy Journal*, *31*(1).

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

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

Al-Shehbaz, I. A., Beilstein, M. A., & Kellogg, E. A. (2006). Systematics and phylogeny of the Brassicaceae (Cruciferae): an overview. *Plant systematics and evolution*, 259, 89-120.

Apel, P., Horstmann, C., & Pfeffer, M. (1997). The *Moricandia* syndrome in species of the Brassicaceae-evolutionary aspects. *Photosynthetica*, *33*, 205-215.

Araus, J. L., Sanchez-Bragado, R., & Vicente, R. (2021). Improving crop yield and resilience through optimization of photosynthesis: panacea or pipe dream?. *Journal of Experimental Botany*, *72*(11), 3936-3955.

Arias, T., & Pires, J. C. (2012). A fully resolved chloroplast phylogeny of the brassica crops and wild relatives (Brassicaceae: Brassiceae): Novel clades and potential taxonomic implications. *Taxon*, 61(5), 980-988.

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.

Arnon, D. I. (1959). Conversion of light into chemical energy in photosynthesis. *Nature*, *184*, 10-21.

Atkinson, R. R., Mockford, E. J., Bennett, C., Christin, P. A., Spriggs, E. L., Freckleton, R. P., ... & Osborne, C. P. (2016). C_4 photosynthesis boosts growth by altering physiology, allocation and size. *Nature plants*, *2*(5), 1-5.

Aubry, S., Brown, N. J., & Hibberd, J. M. (2011). The role of proteins in C_3 plants prior to their recruitment into the C4 pathway. *Journal of experimental botany, 62*(9), 3049-3059.

Bang, S. W., Ueno, O., Wada, Y., Hong, S. K., Kaneko, Y., & Matsuzawa, Y. (2009). Production of *Raphanus sativus* (C_3)-*Moricandia arvensis* (C_3-C_4 intermediate) monosomic and disomic addition lines with each parental cytoplasmic background and their photorespiratory characteristics. *Plant production science*, *12*(1), 70-79.

Barber, J. (2003). Photosystem II: The engine of life. *Quarterly Reviews of Biophysics*, 36(1), 71–89.

Bar-On, Y. M., & Milo, R. (2019). The global mass and average rate of rubisco. *Proceedings of the National Academy of Sciences*, *116*(10), 4738-4743.

Bauwe H., Hagemann M., & Fernie A. R. (2010). Photorespiration: players, partners and origin. *Trends in Plant Science* 15: 330– 336.

Bauwe H., Hagemann M., Kern R., & Timm S. (2012). Photorespiration has a dual origin and manifold links to central metabolism. *Current Opinion in Plant Biology*, 15, 269–275.

Beagan, J. A., & Phillips-Cremins, J. E. (2020). On the existence and functionality of topologically associating domains. *Nature genetics*, 52(1), 8-16.

Bejerano, G., Siepel, A. C., Kent, W. J., & Haussler, D. (2005). Computational screening of conserved genomic DNA in search of functional noncoding elements. *Nature Methods*, *2*(7), 535-545.

Bellasio, C., & Farquhar, G. D. (2019). A leaf-level biochemical model simulating the introduction of C_2 and C_4 photosynthesis in C_3 rice: gains, losses and metabolite fluxes. *New Phytologist*, *223*(1), 150-166.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological), 57*(1), 289-300.

Bernardi, G. (2021). The "genomic code": DNA pervasively molds chromatin structures leaving no room for "Junk". *Life*, *11*(4), 342.

Booker, T. R., Jackson, B. C. & Keightley, P. D. (2017). Detecting positive selection in the genome. *BMC Biol 15*, 98 https://doi.org/10.1186/s12915-017-0434-y

Borghi, G. L., Arrivault, S., Günther, M., Barbosa Medeiros, D., Dell'Aversana, E., Fusco, G. M., ... & Stitt, M. (2022). Metabolic profiles in C_3, C_3-C_4 intermediate, C_4-like, and C_4 species in the genus *Flaveria*. *Journal of experimental botany, 73*(5), 1581-1601.

Bourque, G., Burns, K. H., Gehring, M., Gorbunova, V., Seluanov, A., Hammell, M., ... & Feschotte, C. (2018). Ten things you should know about transposable elements. *Genome biology*, 19, 1-12.

Bowes, G., Ogren, W. L., & Hageman, R. H. (1971). Phosphoglycolate production catalyzed by ribulose diphosphate carboxylase. *Biochemical and Biophysical Research Communications*, 45(3), 716–722. https://doi.org/10.1016/0006-291X(71)90657-1

Bräutigam, A., Kajala, K., Wullenweber, J., Sommer, M., Gagneul, D., Weber, K. L., ... & Weber, A. P. (2011). An mRNA blueprint for C_4 photosynthesis derived from

comparative transcriptomics of closely related C_3 and C_4 species. *Plant physiology*, *155*(1), 142-156.

Bräutigam, A., & Gowik, U. (2016). Photorespiration connects C_3 and C_4 photosynthesis. *Journal of experimental botany*, *67*(10), 2953-2962.

Brocchieri, L. (2001). Phylogenetic inferences from molecular sequences: review and critique. *Theoretical population biology*, 59(1), 27-40.

Broncano, L. S., Pukacz, K. R., Reichel-Deland, V., Schlüter, U., Triesch, S., & Weber, A. P. (2023). Photorespiration is the solution, not the problem. *Journal of Plant Physiology*, *282*, 153928.

Brown R. H., & Morgan J. A. (1980). Photosynthesis of grass species differing in carbon dioxide fixation pathways.:VI. Differential effects of temperature and light intensity on photorespiration in C_3, C_4, and intermediate species. *Plant Physiology 66*, 541–544.

Burgess, S. J., & Hibberd, J. M. (2015). Insights into C_4 metabolism from comparative deep sequencing. *Current Opinion in Plant Biology*, 25, 138-144.

Busch F.A., Sage T.L., Cousins A.B., & Sage R.F. (2013). C_3 plants enhance rates of photosynthesis by reassimilating photorespired and respired CO2. *Plant, Cell & Environment 36*, 200–212.

Busch, F. A., Sage, R. F., & Farquhar, G. D. (2018). Plants increase CO_2 uptake by assimilating nitrogen via the photorespiratory pathway. *Nature plants*, *4*(1), 46-54.

Cai Y., Cao F., Wei K., Zhang G., & Wu F. (2011). Genotypic dependent effect of exogenous glutathione on Cd-induced changes in proteins, ultrastructure and antioxidant defense enzymes in rice seedlings. *Journal of Hazardous Materials*, 192, 1056–1066.

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*, 1-11.

Calvin, M., & Benson, A. A. (1948). The path of carbon in photosynthesis. *Science*, 107(2784), 476–480.

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_4 photosynthesis. *BMC Genomics*, 17, 201.

Cheng, C. Y., Krishnakumar, V., Chan, A. P., Thibaud-Nissen, F., Schobel, S., & Town, C. D. (2017). Araport11: a complete reannotation of the Arabidopsis thaliana reference genome. *The Plant Journal*, 89(4), 789-804.

Clarke, W. E., Higgins, E. E., Plieske, J., Wieseke, R., Sidebottom, C., Khedikar, Y., & Parkin, I. A. (2016). A high-density SNP genotyping array for Brassica napus and its ancestral diploid species based on optimised selection of single-locus markers in the allotetraploid genome. *Theoretical and Applied Genetics*, *129*, 1887-1899.

Cornic, G. (2000). Drought stress inhibits photosynthesis by decreasing stomatal aperture–not by affecting ATP synthesis. *Trends in plant science*, *5*(5), 187-188.

Christin P. A., Besnard G., Samaritani E., Duvall M.R., Hodkinson T.R., Savolainen V., Salamin N. (2008). Oligocene CO_2 decline promoted C_4 photosynthesis in grasses. *Curr Biol* 18:37–43.

Christin, P. A., Osborne, C. P., Sage, R. F., Arakaki, M., & Edwards, E. J. (2011). C_4 eudicots are not younger than C_4 monocots. *Journal of experimental Botany*, 62(9), 3171-3181.

Christin, P. A., & Osborne, C. P. (2014). The evolutionary ecology of C_4 plants. *New Phytologist*, *204*(4), 765-781.

Dahlin, K., König, E., Laubmeier, A., Wehn, A., & Rıos-Soto, K. (2012). Competition Model between the Invasive Sahara Mustard and Native Plants in the Sonoran Desert. *Society for Advancement of Chicanos/Hispanics and Native Americans in Science, Santa Cruz, CA*.

Dal'Molin, C. G. O., Quek, L. E., Palfreyman, R. W., Brumbley, S. M., & Nielsen, L. K. (2010). C4GEM, a genome-scale metabolic model to study C4 plant metabolism. *Plant physiology*, *154*(4), 1871-1885.

Darst, R. P., Pardo, C. E., Ai, L., Brown, K. D., & Kladde, M. P. (2010). Bisulfite sequencing of DNA. Current protocols in molecular biology, 91(1), 7-9.

Ding, Z., Weissmann, S., Wang, M., Du, B., Huang, L., Wang, L., ... & Li, P. (2015). Identification of photosynthesis-associated C_4 candidate genes through comparative leaf gradient transcriptome in multiple lineages of C3 and C4 species. *PloS one*, *10*(10), e0140629.

Dong, P., Tu, X., Li, H., Zhang, J., Grierson, D., Li, P., & Zhong, S. (2020). Tissue-specific Hi-C analyses of rice, foxtail millet and maize suggest non-canonical function of plant chromatin domains. *Journal of Integrative Plant Biology*, *62*(2), 201-217.

Doring F., Streubel M., Brautigam A., Gowik U. (2016) Most photorespiratory genes are preferentially expressed in the bundle sheath cells of the C_4 grass Sorghum bicolor. *Journal of Experimental Botany, 67*, 3053–3064.

Deaton, A. M., & Bird, A. (2011). CpG islands and the regulation of transcription. *Genes & development*, 25(10), 1010-1022.

Degnan, J. H., & Rosenberg, N. A. (2009). Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in ecology & evolution*, 24(6), 332-340.

Duarte-Aké, F., Us-Camas, R., Cancino-García, V. J., & De-la-Peña, C. (2019). Epigenetic changes and photosynthetic plasticity in response to environment. *Environmental and Experimental Botany*, *159*, 108-120.

Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association, 56*(293), 52-64.

de-Leon, S. B. T., & Davidson, E. H. (2007). Gene regulation: gene control network in development. *Annu. Rev. Biophys. Biomol. Struct*., 36, 191-212.

Edwards, E. J., Ku, M. B., Hatch, M. D., & Boardman, N. K. (1987). The biochemistry of plants: a comprehensive treatise, Vol. 14: Photosynthesis.

Edwards, G. E., & Voznesenskaya, E. V. (2011). Chapter 4 C_4 photosynthesis: Kranz forms and single-cell C_4 in terrestrial plants. C_4 photosynthesis and related CO2 concentrating mechanisms, 29-61.

Edwards, E. J. (2014). The inevitability of C_4 photosynthesis. *elife*, *3*, e03702.

Edwards E. J. (2019). Evolutionary trajectories, accessibility and other metaphors: the case of C_4 and CAM photosynthesis. *New Phytologist 223*: 1742– 1755.

Eisenhut, M., Roell, M. S., & Weber, A. P. (2019). Mechanistic understanding of photorespiration paves the way to a new green revolution. *New Phytologist*, *223*(4), 1762-1769.

El Hasnaoui, S., Fahr, M., & Smouni, A. (2024). Genome architecture of the heavy metal tolerant and accumulator *Hirschfeldia incana*: Insights from genome sequencing, assembly, and comparative analysis. *Environmental and Experimental Botany*, *228*, 105991.

Emms, D. M., & Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome biology*, *20*, 1-14.

Emrich, S. J., Aluru, S., Fu, Y., Wen, T. J., Narayanan, M., Guo, L., ... & Schnable, P. S. (2004). A strategy for assembling the maize (Zea mays L.) genome. *Bioinformatics*, *20*(2), 140-147.

Evans, T. G. (2015). Considerations for the use of transcriptomics in identifying the 'genes that matter' for environmental adaptation. *The Journal of experimental biology, 218*(12), 1925-1935.

Farmer, A., Thibivilliers, S., Ryu, K. H., Schiefelbein, J., & Libault, M. (2021). Single-nucleus RNA and ATAC sequencing reveals the impact of chromatin accessibility on gene expression in Arabidopsis roots at the single-cell level. *Molecular Plant, 14*(3), 372-383.

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

Flügel F, Timm S, Arrivault S, Florian A, Stitt M, Fernie AR, & Bauwe H. (2017). The Photorespiratory Metabolite 2-Phosphoglycolate Regulates Photosynthesis and Starch Accumulation in *Arabidopsis*. *The Plant Cell* 29, 2537–2551.

Freire, R., Weisweiler, M., Guerreiro, R., Baig, N., Hüttel, B., Obeng-Hinneh, E., ... & Stich, B. (2021). Chromosome-scale reference genome assembly of a diploid potato clone derived from an elite variety. *G3*, *11*(12), jkab330.

Foyer, C. H., Bloom, A. J., Queval, G., & Noctor, G. (2009). Photorespiratory metabolism: genes, mutants, energetics, and redox signaling. *Annual review of plant biology*, *60*(1), 455-484.

Furbank, R. T., & Taylor, W. C. (1995). Regulation of photosynthesis in C_3 and C_4 plants: a molecular approach. *The plant cell*, *7*(7), 797.

Futuyma, D. J., & Moreno, G. (1988). The evolution of ecological specialization. *Annual Review of Ecology and Systematics, 19*(1), 207–233.

Gabaldón, T., & Koonin, E. V. (2013). Functional and evolutionary implications of gene orthology. *Nature Reviews Genetics*, *14*(5), 360-366.

Gabur, I., Chawla, H. S., Snowdon, R. J., & Parkin, I. A. (2019). Connecting genome structural variation with complex traits in crop plants. *Theoretical and applied genetics*, *132*, 733-750.

Ganal M. W., Altmann T., & Roder M.S. (2009). SNP identification in crop plants. *Curr Opin Plant Biol* 12:211–217.

Ganal M.W., Polley A., Graner E. M., Plieske J., Wieseke R., Luerssen H. & Durstewitz G. (2012). Large SNP arrays for genotyping in crop plants. *J Biosci* 37:821–828.

Gevaert, K., Impens, F., Ghesquière, B., Van Damme, P., Lambrechts, A., & Vandekerckhove, J. (2008). Stable isotopic labeling in proteomics. *Proteomics*, *8*(23-24), 4873-4885.

Ghannoum, O. (2009). C_4 photosynthesis and water stress. *Annals of botany*, *103*(4), 635-644.

Ghannoum, O. & von Caemmerer, S. (2011). Nitrogen and Water Use Efficiency of C. *Concentrating Mechanisms*, 129.

Gioria, M., Hulme, P. E., Richardson, D. M., & Pyšek, P. (2023). Why are invasive plants successful?. *Annual Review of Plant Biology*, *74*(1), 635-670.

Gomez J.M., Perfetti F, Armas C., Narbona E., Gonzalez-Megias A., Navarro L., DeSoto L. & Torices R. (2020). Within-individual phenotypic plasticity in flowers fosters pollination niche shift. *Nature Communications* 11, 4019.

Gorai, M., Tlig, T., & Neffati, M. (2009). Influence of water stress on seed germination characteristics in invasive *Diplotaxis harra* (Forssk.) Boiss (Brassicaceae) in arid zone of Tunisia. *Journal of phytology*, *1*(4).

Gowik, U., & Westhoff, P. (2011). The path from C_3 to C_4 photosynthesis. *Plant physiology*, *155*(1), 56-63.

Grass Phylogeny Working Group II. (2012). New grass phylogeny resolves deep evolutionary relationships and discovers C_4 origins. *New Phytologist*, 193(2), 304-312.

Griffiths, H., Weller, G., Toy, L. F., & Dennis, R. J. (2013). You're so vein: bundle sheath physiology, phylogeny and evolution in C_3 and C_4 plants. *Plant, cell & environment*, *36*(2), 249-261.

Guan X.Q., Zhao S.J., Li D.Q., & Shu H.R. (2004). Photoprotective function of photorespiration in several grapevine cultivars under drought stress. *Photosynthetica*, 42, 31–36.

Guo Z., Ou W., Lu S. & Zhong Q. (2006). Differential responses of antioxidative system to chilling and drought in four rice cultivars differing in sensitivity. *Plant Physiology and Biochemistry*, 44, 828–836.

Gramazio, P., Prohens, J., Toppino, L., & Plazas, M. (2021). Introgression breeding in cultivated plants. *Frontiers in Plant Science*, *12*, 764533.

Hassan, A. H., Mokhtar, M. M., & El Allali, A. (2024). Transposable elements: multifunctional players in the plant genome. *Frontiers in Plant Science, 14*, 1330127.

Hatch, M. D., C. R. Slack, & H. S. Johnson. (1967). Further studies on a new pathway of photosynthetic carbon dioxide fixation in sugarcane and its occurrence in other plant species. *Biochem. J.* 102: 417-42

Hatch, M. D. (1987). C_4 photosynthesis: a unique elend of modified biochemistry, anatomy and ultrastructure. *Biochimica et Biophysica Acta (BBA)-Reviews on Bioenergetics*, *895*(2), 81-106.

Haupt-Herting S., & Fock H.P. (2002). Oxygen exchange in relation to carbon assimilation in water-stressed leaves during photosynthesis. *Annals of Botany*, 89, 851–859.

Heckmann, D., Schulze, S., Denton, A., Gowik, U., Westhoff, P., Weber, A. P., & Lercher, M. J. (2013). Predicting C_4 photosynthesis evolution: modular, individually adaptive steps on a Mount Fuji fitness landscape. *Cell*, *153*(7), 1579-1588.

Hedrick, P. W. (2007). Balancing selection. *Current Biology*, *17*(7), R230-R231.

Hiller, M., Schaar, B.T., Indjeian, V.B., Kingsley, D.M., Hagey, L.R. & Bejerano, G. (2012). A “forward genomics” approach links genotype to phenotype using independent phenotypic losses among related species. *Cell Reports, 2*(4), 817–823.

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., Walden, N., Caracciolo, L., Luoni, S. B., Retta, M., Li, R., ... & Schranz, M. E. (2024). Expanding the Triangle of U: The genome assembly of *Hirschfeldia incana* provides insights into chromosomal evolution, phylogenomics and high photosynthesis-related traits. *bioRxiv*, 2024-05.

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

Igamberdiev A.U., Bykova N.V., Lea P.J., & Gardeström P. (2001). The role of photorespiration in redox and energy balance of photosynthetic plant cells: a study with a barley mutant deficient in glycine decarboxylase. *Physiologia Plantarum*, 111, 427–438.

Jaganathan, K., Panagiotopoulou, S. K., McRae, J. F., Darbandi, S. F., Knowles, D., Li, Y. I. & Farh, K. K. H. (2019). Predicting splicing from primary sequence with deep learning. *Cell*, 176(3), 535-548.

Jay, Jethva., F., Hahn., Rita, Giuliani., Niels, Peeters., Asaph, B., Cousins., Steven, Kelly. (2024). Realisation of a key step in the evolution of C_4 photosynthesis in rice by genome editing. bioRxiv, doi: 10.1101/2024.05.21.595093

Johnson, A. D., Handsaker, R. E., Pulit, S. L., Nizzari, M. M., O'Donnell, C. J., & De Bakker, P. I. (2008). SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics*, 24(24), 2938-2939.

Johnson, J. E., Field, C. B., & Berry, J. A. (2021). The limiting factors and regulatory processes that control the environmental responses of C_3, C_3–C_4 intermediate, and C_4 photosynthesis. *Oecologia*, 1-26.

Julkowska, M.M. (2024). Data insights: descriptive data papers with high information impact are a new gold mine for readers of plant, cell and environment. *Plant, Cell and Environment*, 1–2. https://doi.org/10.1111/pce.14804

Leegood, R. C., & von Caemmerer, S. (1994). Regulation of photosynthetic carbon assimilation in leaves of C_3–C_4 intermediate species of *Moricandia* and *Flaveria*. *Planta, 192,* 232-238.

Li, Y., Heckmann, D., Lercher, M. J., & Maurino, V. G. (2017). Combining genetic and evolutionary engineering to establish C_4 metabolism in C_3 plants. *Journal of Experimental Botany*, *68*(2), 117-125.

Li, Y. T., Luo, J., Liu, P., & Zhang, Z. S. (2021). C_4 species utilize fluctuating light less efficiently than C_3 species. *Plant Physiology*, *187*(3), 1288-1291.

Li, X., Wang, J., Yu, Y., Li, G., Wang, J., Li, C., ... & Gong, L. (2023). Genomic rearrangements and evolutionary changes in 3D chromatin topologies in the cotton tribe (Gossypieae). *BMC biology, 21*(1), 56.

Liao, I. T., Sears, K. E., Hileman, L. C., & Nikolov, L. A. (2024). Different orthology inference algorithms generate similar predicted orthogroups among Brassicaceae species. *Applications in Plant Sciences*, e11627.

Lieberman-Aiden, E., Van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., ... & Dekker, J. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, *326*(5950), 289-293.

Liu, Q., Teng, S., Deng, C., Wu, S., Li, H., Wang, Y., ... & Lu, T. (2023). SHORT ROOT and INDETERMINATE DOMAIN family members govern PIN-FORMED expression to regulate minor vein differentiation in rice. *The Plant Cell*, koad125.

Long, H. K., Prescott, S. L., & Wysocka, J. (2016). Ever-changing landscapes: transcriptional enhancers in development and evolution. Cell, 167(5), 1170-1187.

Lundgren, M. R., & Christin, P. A. (2017). Despite phylogenetic effects, C_3–C_4 lineages bridge the ecological gap to C4 photosynthesis. Journal of Experimental Botany, 68(2), 241-254.

Lundgren, M. R. (2020). C_2 photosynthesis: a promising route towards crop improvement? *New Phytologist*, 228(6), 1734-1740.

Kadereit, G., Bohley, K., Lauterbach, M., Tefarikis, D. T., & Kadereit, J. W. (2017). C_3–C_4 intermediates may be of hybrid origin–a reminder. *New Phytologist*, 215(1), 70-76.

Khoshravesh, R., Stinson, C. R., Stata, M., Busch, F. A., Sage, R. F., Ludwig, M., & Sage, T. L. (2016). C_3–C_4 intermediacy in grasses: organelle enrichment and distribution, glycine decarboxylase expression, and the rise of C_2 photosynthesis. *Journal of Experimental Botany*, 67(10), 3065-3078.

Kiefer, C., Willing, E. M., Jiao, W. B., Sun, H., Piednoël, M., Hümann, U., ... & Schneeberger, K. (2019). Interspecies association mapping links reduced CG to TG substitution rates to the loss of gene-body methylation. *Nature Plants, 5*(8), 846-855.

Kim, D., & Sung, S. (2017). The regulatory switch of Polycomb Repressive Complexes in plants. *Plant Signal Behav.*, 12(1), e1284726.

Kikuchi, G., & Hiraga, K. (1982). The mitochondrial glycine cleavage system: unique features of the glycine decarboxylation. *Molecular and cellular biochemistry*, *45*, 137-149.

Klasberg, S., Bitard-Feildel, T., & Mallet, L. (2016). Computational identification of novel genes: current and future perspectives. *Bioinformatics and Biology insights*, *10*, BBI-S39950.

Krenzer Jr, E. G., Moss, D. N., & Crookston, R. K. (1975). Carbon dioxide compensation points

of flowering plants. *Plant Physiology, 56*(2), 194-206.

Kress, W. J., Soltis, D. E., Kersey, P. J., Wegrzyn, J. L., Leebens-Mack, J. H., Gostel, M. R., Liu, X. & Soltis, P. S. (2022). Green plant genomes: What we know in an era of rapidly expanding opportunities. *Proceedings of the National Academy of Sciences*, *119*(4), e2115640118.

Ku, M. S., Monson, R. K., Littlejohn Jr, R. O., Nakamoto, H., Fisher, D. B., & Edwards, G. E. (1983). Photosynthetic characteristics of C_3-C_4 intermediate *Flaveria* species: I. Leaf anatomy, photosynthetic responses to O_2 and CO_2, and activities of key enzymes in the C_3 and C_4 pathways. *Plant physiology*, *71*(4), 944-948.

Ku, M. S., Wu, J., Dai, Z., Scott, R. A., Chu, C., & Edwards, G. E. (1991). Photosynthetic and photorespiratory characteristics of *Flaveria* species. *Plant physiology*, 96(2), 518-528.

Mabry, M. E., Abrahams, R. S., Al-Shehbaz, I. A., Baker, W. J., Barak, S., Barker, M. S., & Harkess, A. E. (2023). Complementing model species with model clades. *The Plant Cell*, koad260.

Malhi, G. S., Kaur, M., & Kaushik, P. (2021). Impact of climate change on agriculture and its mitigation strategies: A review. *Sustainability*, *13*(3), 1318.

Makino, A. (2011). Photosynthesis, grain yield, and nitrogen utilization in rice and wheat. *Plant physiology*, *155*(1), 125-129.

Mardis, E. R. (2007). ChIP-seq: welcome to the new frontier. *Nature methods*, *4*(8), 613-614.

Massingham, T., & Goldman, N. (2005). Detecting amino acid sites under positive selection and purifying selection. *Genetics*, 169(3), 1753-1762.

McCarthy I., Romero-Puertas M.C., Palma J.M., Sandalio L.M., Corpas F.J., Gómez M., & del Rio L.A. (2001). Cadmium induces senescence symptoms in leaf peroxisomes of pea plants. *Plant, Cell and Environment*, 24, 1065–1073.

McCarroll, S. A., & Altshuler, D. M. (2007). Copy-number variation and association studies of human disease. *Nature genetics*, 39(Suppl 7), S37-S42.

McCormick, R. F., Truong, S. K., Sreedasyam, A., Jenkins, J., Shu, S., Sims, D., ... & Mullet, J. E. (2018). The *Sorghum bicolor* reference genome: improved assembly, gene annotations, a transcriptome atlas, and signatures of genome organization. *The Plant Journal*, 93(2), 338-354.

McKown, A. D., Moncalvo, J. M., & Dengler, N. G. (2005). Phylogeny of *Flaveria* (Asteraceae) and inference of C_4 photosynthesis evolution. *American Journal of Botany*, 92(11), 1911-1928.

McKown, A. D., & Dengler, N. G. (2007). Key innovations in the evolution of Kranz anatomy and C_4 vein pattern in *Flaveria* (Asteraceae). *American journal of botany, 94*(3), 382-399.

Mercado, M. A., & Studer, A. J. (2022). Meeting in the Middle: Lessons and Opportunities from Studying C_3-C_4 Intermediates. *Annual Review of Plant Biology*, 73, 43-65.

Meng, F., Zhao, H., Zhu, B., Zhang, T., Yang, M., Li, Y., ... & Jiang, J. (2021). Genomic editing of intronic enhancers unveils their role in fine-tuning tissue-specific gene expression in *Arabidopsis thaliana*. *The Plant Cell*, 33(6), 1997-2014.

Michael, T. P., & Jackson, S. (2013). The first 50 plant genomes. *Plant Genome*, *6*(2), 1-7.

Miglani, G. S., Kaur, R., Sharma, P., & Gupta, N. (2021). Leveraging photosynthetic efficiency toward improving crop yields. *Journal of Crop Improvement*, 35(3), 361-402.

Mitchell, P. (1961). Coupling of phosphorylation to electron and hydrogen transfer by a chemiosmotic type of mechanism. *Nature*, 191(4784), 144–148.

Mhiri, C., Borges, F., & Grandbastien, M. A. (2022). Specificities and dynamics of transposable elements in land plants. *Biology, 11*(4), 488.

Miki, D., Zinta, G., Zhang, W., Peng, F., Feng, Z., & Zhu, J. K. (2021). CRISPR/Cas9-based genome editing toolbox for *Arabidopsis thaliana*. *Arabidopsis Protocols*, 121-146.

Mittova V., Tal M., Volokita M., Guy M. (2003). Up-regulation of the leaf mitochondrial and peroxisomal antioxidative systems in response to salt-induced oxidative stress in the wild salt-tolerant tomato species *Lycopersicon pennellii*. *Plant, Cell and Environment*, 26, 845–856.

Monson, R. K., Edwards, G. E., & Ku, M. S. (1984). C_3-C_4 intermediate photosynthesis in plants. *Bioscience*, *34*(9), 563-574.

Monson, R. K. (1989). The relative contributions of reduced photorespiration, and improved water-and nitrogen-use efficiencies, to the advantages of C_3-C_4 intermediate photosynthesis in *Flaveria*. *Oecologia*, *80*, 215-221.

Monson, R. K., & Rawsthorne, S. (2000). CO_2 assimilation in C_3-C_4 intermediate plants. Photosynthesis: physiology and metabolism, 533-550.

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

Mostovoy, Y., Levy-Sakin, M., Lam, J., Lam, E. T., Hastie, A. R., Marks, P., ... & Kwok, P. Y. (2016). A hybrid approach for *de novo* human genome sequence assembly and phasing. *Nature methods*, 13(7), 587-590.

Muhaidat, R., Sage, R. F., & Dengler, N. G. (2007). Diversity of Kranz anatomy and biochemistry in C_4 eudicots. *American journal of botany*, 94(3), 362-381.

Muhaidat, R., Sage, T. L., Frohlich, M. W., Dengler, N. G., & Sage, R. F. (2011). Characterization of C_3–C_4 intermediate species in the genus *Heliotropium* L.(Boraginaceae): anatomy, ultrastructure and enzyme activity. *Plant, Cell & Environment*, 34(10), 1723-1736.

Munekage, Y. N., & Taniguchi, Y. Y. (2022). A scheme for C4 evolution derived from a comparative analysis of the closely related C_3, C_3–C_4 intermediate, C_4-like, and C_4 species in the genus *Flaveria*. *Plant Molecular Biology, 110*(4), 445-454.

Nagy, L. G., Ohm, R. A., Kovács, G. M., Floudas, D., Riley, R., Gácser, A., ... & Hibbett, D. S. (2014). Latent homology and convergent regulatory evolution underlies the repeated emergence of yeasts. *Nature communications*, 5(1), 4471.

Nagy, L.G., Merényi, Z., Hegedüs, B. & Bálint, B. (2020). Novel phylogenetic methods are needed for understanding gene function in the era of mega-scale genome sequencing. *Nucleic Acids Research*, 48(5), 2209–2219.

Nawy, T. (2014). Single-cell sequencing. *Nature methods*, *11*(1), 18-18.

Nikolov, L. A., Shushkov, P., Nevado, B., Gan, X., Al-Shehbaz, I. A., Filatov, D., ... & Tsiantis, M. (2019). Resolving the backbone of the Brassicaceae phylogeny for investigating trait diversity. *New Phytologist*, 222(3), 1638-1651.

Lin, M. Y., Schlüter, U., Stich, B., & Weber, A. P. (2021). Cis-regulatory divergence underpins the evolution of C_3-C_4 intermediate photosynthesis in *Moricandia*. *bioRxiv*, 2021-05.

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

Parry, M. A. J., Andralojc, P. J., Mitchell, R. A., Madgwick, P. J., & Keys, A. J. (2003). Manipulation of Rubisco: the amount, activity, function and regulation. *Journal of experimental botany*, *54*(386), 1321-1333.

Pérez-Soba, M., Petit, S., Jones, L., Bertrand, N., Briquel, V., Omodei-Zorini, L., ... & De Groot, R. (2008). Land use functions—A multifunctionality approach to assess the impact of land use changes on land use sustainability. *Sustainability impact assessment of land use changes*, 375-404.

Pignon, C. P., Jaiswal, D., McGrath, J. M., & Long, S. P. (2017). Loss of photosynthetic efficiency in the shade. An Achilles heel for the dense modern stands of our most productive C_4 crops? *Journal of Experimental Botany*, *68*(2), 335-345.

Pignon, C. P., Lundgren, M. R., Osborne, C. P., & Long, S. P. (2019). Bundle sheath chloroplast volume can house sufficient Rubisco to avoid limiting C_4 photosynthesis during chilling. *Journal of experimental botany*, *70*(1), 357-365.

Pinheiro, C., Emiliani, G., Marino, G., Fortunato, A. S., Haworth, M., De Carlo, A., ... & Centritto, M. (2023). Metabolic background, not photosynthetic physiology, determines drought and drought recovery responses in C_3 and C_2 Moricandias. *International Journal of Molecular Sciences*, *24*(4), 4094.

Pollard, M. O., Gurdasani, D., Mentzer, A. J., Porter, T., & Sandhu, M. S. (2018). Long reads: their purpose and place. *Human molecular genetics*, *27*(R2), R234-R241.

Prudent, X., Parra, G., Schwede, P., Roscito, J.G. & Hiller, M. (2016). Controlling for phylogenetic relatedness and evolutionary rates improves the discovery of associations between species' phenotypic and genomic differences. *Molecular Biology and Evolution, 33*(8), 2135–2150.

Quail, M. A., Kozarewa, I., Smith, F., Scally, A., Stephens, P. J., Durbin, R., ... & Turner, D. J. (2008). A large genome center's improvements to the Illumina sequencing system. *Nature methods*, 5(12), 1005-1010.

Raines, C. A. (2011). Increasing photosynthetic carbon assimilation in C_3 plants to improve crop yield: current and future strategies. *Plant physiology*, *155*(1), 36-42.

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 C_3-C_4 intermediate species of *Moricandia*. *Planta*, *173*, 298-308.

Rawsthorne, S., Hylton, C. M., Smith, A. M., & Woolhouse, H. W. (1988). Distribution of photorespiratory enzymes between bundle-sheath and mesophyll cells in leaves of the C_3− C_4 intermediate species *Moricandia arvensis* (L.) DC. *Planta*, *176*, 527-532.

Rawsthorne S. (1992). C3–C4 intermediate photosynthesis: linking physiology to gene expression. *The Plant Journal* **2**, 267–274.

Rishmawi, L., Bühler, J., Jaegle, B., Hülskamp, M., & Koornneef, M. (2017). Quantitative trait loci controlling leaf venation in Arabidopsis. *Plant, Cell and Environment*, 40(8), 1429-1441.

Rao, A., Barkley, D., França, G. S., & Yanai, I. (2021). Exploring tissue architecture using spatial transcriptomics. *Nature*, *596*(7871), 211-220.

Rodermel, S. (1999). Subunit control of Rubisco biosynthesis–a relic of an endosymbiotic past?. *Photosynthesis Research*, *59*, 105-123.

Romero-Puertas M.C., Palma J.M., Gómez M., Del Río L.A., & Sandalio L.M. (2002). Cadmium causes the oxidative modification of proteins in pea plants. *Plant, Cell and Environment*, 25, 677–686.

Rylott, E. L., Metzlaff, K., & Rawsthorne, S. (1998). Developmental and environmental effects on the expression of the C_3-C_4 intermediate phenotype in *Moricandia arvensis*. *Plant physiology, 118*(4), 1277-1284.

Salvucci, M. E., & Crafts-Brandner, S. J. (2004). Inhibition of photosynthesis by heat stress: the activation state of Rubisco as a limiting factor in photosynthesis. *Physiologia plantarum*, *120*(2), 179-186.

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

Sage, R. F., & McKown, A. D. (2006). Is C4 photosynthesis less phenotypically plastic than C3 photosynthesis? *Journal of experimental botany*, *57*(2), 303-317.

Sage R. F., Christin P. A., Edwards E. J. (2011). The C4 plant lineages of planet Earth. *Journal of Experimental Botany* **62**, 3155–3169

Sage, R. F., Sage, T. L., & Kocacinar, F. (2012). Photorespiration and the evolution of C4 photosynthesis. *Annual review of plant biology*, *63*(1), 19-47.

Sage, R. F., Khoshravesh, R., & Sage, T. L. (2014). From proto-Kranz to C_4 Kranz: building the bridge to C4 photosynthesis. *Journal of experimental botany*, *65*(13), 3341-3356.

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

Sampaio, B. L., Edrada-Ebel, R., & Da Costa, F. B. (2016). Effect of the environment on the secondary metabolic profile of *Tithonia diversifolia*: a model for environmental metabolomics of plants. *Scientific reports, 6*(1), 29265.

Schjoerring, J. K., Husted, S., Mäck, G., Nielsen, K. H., Finnemann, J., & Mattsson, M. (2000). Physiological regulation of plant-atmosphere ammonia exchange. *Plant and soil*, 221, 95-102.

Schlüter, U., & Weber, A. P. (2016). The road to C_4 photosynthesis: evolution of a complex trait via intermediary states. *Plant and Cell Physiology*, *57*(5), 881-889.

Schlüter, U., Bräutigam, A., Gowik, U., Melzer, M., Christin, P. A., Kurz, S., Mettler-Altmann, T. & Weber, A. P. (2017). Photosynthesis in C_3–C_4 intermediate *Moricandia* species. *Journal of Experimental Botany*, *68*(2), 191-206.

Schuster, W. S., & Monson, R. K. (1990). An examination of the advantages of C_3-C_4 intermediate photosynthesis in warm environments. *Plant, Cell & Environment, 13*(9), 903-912.

Schuler, M. L., Mantegazza, O., & Weber, A. P. (2016). Engineering C_4 photosynthesis into C_3 chassis in the synthetic biology age. *The Plant Journal*, *87*(1), 51-65.

Shlyueva, D., Stampfel, G., & Stark, A. (2014). Transcriptional enhancers: from properties to genome-wide predictions. *Nature Reviews Genetics*, 15(4), 272-286.

Schmidt, M., Guerreiro, R., Baig, N., Habekuß, A., Will, T., Ruckwied, B., & Stich, B. (2024). Fine mapping a QTL for BYDV-PAV resistance in maize. *In review*

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

Sedelnikova, O. V., Hughes, T. E., & Langdale, J. A. (2018). Understanding the genetic basis of C4 Kranz anatomy with a view to engineering C3 crops. *Annual review of genetics*, *52*(1), 249-270.

Singh, P., & Reeves, G. (2020). Constructing the bundle sheath towards enhanced photosynthesis. *Journal of Experimental Botany*, *71*(4), 1206-1209.

Smith, E. N., van Aalst, M., Tosens, T., Niinemets, Ü., Stich, B., Morosinotto, T., ... & Weber, A. P. (2023). Improving photosynthetic efficiency toward food security: Strategies, advances, and perspectives. *Molecular Plant.*

Smith, S.D., Pennell, M.W., Dunn, C.W. & Edwards, S.V. (2020). Phylogenetics is the new genetics (for most of biodiversity). *Trends in Ecology & Evolution, 35*(5), 415–425.

South, P. F., Cavanagh, A. P., Liu, H. W., & Ort, D. R. (2019). Synthetic glycolate metabolism pathways stimulate crop growth and productivity in the field. *Science*, *363*(6422), eaat9077.

Stanke, M., & Waack, S. (2003). Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics-Oxford*, 19(2), 215-225.

Stata, M., Sage, T. L., Hoffmann, N., Covshoff, S., Ka-Shu Wong, G., & Sage, R. F. (2016). Mesophyll chloroplast investment in C_3, C_4 and C_2 species of the genus Flaveria. *Plant and Cell Physiology*, 57(5), 904-918.

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*, 36(22-23), 5291-5298.

Taniguchi, Y., Ohkawa, H., Masumoto, C., Fukuda, T., Tamai, T., Lee, K., ... & Miyao, M. (2008). Overproduction of C_4 photosynthetic enzymes in transgenic rice plants: an approach to introduce the C_4-like photosynthetic pathway into rice. *Journal of experimental botany*, 59(7), 1799-1809.

Taylor, G., Garassino, F., Aarts, M. G., & Harbinson, J. (2023). Improving C_3 photosynthesis by exploiting natural genetic variation: *Hirschfeldia incana* as a model species. *Food and Energy Security*, *12*(1), e420.

Timm, S., Mielewczik, M., Florian, A., Frankenbach, S., Dreissen, A., Hocken, N., ... & Bauwe, H. (2012). High-to-low CO_2 acclimation reveals plasticity of the photorespiratory pathway and indicates regulatory links to cellular metabolism of *Arabidopsis*. *PLoS One 7*(8), e42809.

Timm S., & Bauwe H. (2013). The variety of photorespiratory phenotypes—employing the current status for future research directions on photorespiration. *Plant Biology 15*, 737–747.

Thrash, A., Hoffmann, F., & Perkins, A. (2020). Toward a more holistic method of genome assembly assessment. *BMC bioinformatics*, *21*(Suppl 4), 249.

Treangen, T. J., & Salzberg, S. L. (2012). Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics*, *13*(1), 36-46.

Trifinopoulos, J., Nguyen, L. T., von Haeseler, A. & Minh, B.Q. (2016) WIQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. *Nucleic Acids Research, 44*(W1), W232–W235.

Walker, B. J., South, P. F., & Ort, D. R. (2016). Physiological evidence for plasticity in glycolate/glycerate transport during photorespiration. *Photosynthesis Research*, *129*(1),

93-103.

Walsh, C. A. (2023). Bringing home the carbon: photorespiratory CO_2 recovery shows diverse efficiency in Brassicaceae. *Journal of Experimental Botany*, 74(21), 6399-6404.

Walsh, C. A., Bräutigam, A., Roberts, M. R., & Lundgren, M. R. (2023). Evolutionary implications of C_2 photosynthesis: how complex biochemical trade-offs may limit C_4 evolution. *Journal of Experimental Botany*, *74*(3), 707-722.

Wang, X., Gowik, U., & Tang, H. (2013). Genome-scale metabolic network of the C4 plant Flaveria bidentis. *Frontiers in Plant Science*, 4, 62.

Wang, P., Khoshravesh, R., Karki, S., Tapia, R., Balahadia, C. P., Bandyopadhyay, A., ... & Langdale, J. A. (2017). Re-creation of a key step in the evolutionary switch from C_3 to C_4 leaf anatomy. *Current Biology*, *27*(21), 3278-3287.

Wang, R., Wang, Z., Wang, J., & Li, S. (2019). SpliceFinder: ab initio prediction of splice sites using convolutional neural network. *BMC bioinformatics*, 20, 1-13.

Wang, L., Ma, G., Wang, H., Cheng, C., Mu, S., Quan, W. & Zhang, Y. (2019). A draft genome assembly of halophyte Suaeda aralocaspica, a plant that performs C4 photosynthesis within individual cells. *GigaScience*, 8(9), giz116.

Warwick, S. I., & Black, L. D. (1991). Molecular systematics of Brassica and allied genera (subtribe Brassicinae, Brassiceae)—chloroplast genome and cytodeme congruence. *Theoretical and Applied Genetics*, 82, 81-92.

Warwick, S. I., & Hall, J. C. (2009). Phylogeny of Brassica and wild relatives. *Biology and breeding of crucifers,* 19, 36.

Weber A.P. (2015). Discovering new biology through RNA-Seq. *Plant Physiol*.169: 1524–1531.

Weissbach, A., Horecker, B. L., & Hurwitz, J. (1956). The enzymatic formation of phosphoglyceric acid from ribulose diphosphate and carbon dioxide. *Journal of*

Biological Chemistry, *218*(2), 795-810.

White, J. K., Gerdin, A. K., Karp, N. A., Ryder, E., Buljan, M., Bussell, J. N. & Steel, K. P. (2013). Genome-wide generation and systematic phenotyping of knockout mice reveals new roles for many genes. *Cell*, *154*(2), 452-464.

Wingler, A., Lea, P. J., Quick, W. P., & Leegood, R. C. (2000). Photorespiration: metabolic pathways and their role in stress protection. Philosophical Transactions of the Royal Society of London. Series B: *Biological Sciences*, 355(1402), 1517-1529.

Winkler, D. E., Chapin, K. J., François, O., Garmon, J. D., Gaut, B. S., & Huxman, T. E. (2019). Multiple introductions and population structure during the rapid expansion of the invasive Sahara mustard (*Brassica tournefortii*). *Ecology and Evolution*, *9*(14), 7928-7941.

Ueno, O., Bang, S. W., Wada, Y., Kondo, A., Ishihara, K., Kaneko, Y., & Matsuzawa, Y. (2003). Structural and biochemical dissection of photorespiration in hybrids differing in genome constitution between Diplotaxis tenuifolia (C_3-C_4) and radish (C_3). *Plant Physiology,* 132(3), 1550-1559.

Ueno, O., Wada, Y., Wakai, M., & Bang, S. W. (2006). Evidence from photosynthetic characteristics for the hybrid origin of *Diplotaxis muralis* from a C_3-C_4 intermediate and a C_3 species. *Plant Biology*, *8*(02), 253-259.

Ueno, O., Bang, S. W., Wada, Y., Kobayashi, N., Kaneko, R., Kaneko, Y., & Matsuzawa, Y. (2007). Inheritance of C_3-C_4 intermediate photosynthesis in reciprocal hybrids between *Moricandia arvensis* (C_3-C_4) and *Brassica oleracea* (C_3) that differ in their genome constitution. *Plant production science, 10*(1), 68-79.

Ueno O. (2011). Structural and biochemical characterization of the C_3-C_4 intermediate *Brassica gravinae* and relatives, with particular reference to cellular distribution of Rubisco. *Journal of Experimental Botany 62*, 5347–5355.

Ullah, F., Hamilton, M., Reddy, A. S., & Ben-Hur, A. (2018). Exploring the relationship between intron retention and chromatin accessibility in plants. *BMC genomics*, *19*(1), 1-11.

van Rooijen, R., Schulze, S., Petzsch, P., & Westhoff, P. (2020). Targeted misexpression of NAC052, acting in H3K4 demethylation, alters leaf morphological and anatomical traits in *Arabidopsis thaliana*. *Journal of Experimental Botany*, *71*(4), 1434-1448.

Vergara, Z., & Gutierrez, C. (2017). Emerging roles of chromatin in the maintenance of genome organization and function in plants. *Genome Biology*, *18*(1), 1-12.

Vogan P.J., Frohlich M.W. & Sage R.F. (2007) The functional significance of C_3-C_4 intermediate traits in *Heliotropium* L. (Boraginaceae): gas exchange perspectives. *Plant, Cell & Environment* 30, 1337–1345.

Vogan, P. J., & Sage, R. F. (2011). Water-use efficiency and nitrogen-use efficiency of C_3-C_4 intermediate species of *Flaveria* Juss.(Asteraceae). *Plant, cell & environment*, *34*(9), 1415-1430.

Voss, I., Sunil, B., Scheibe, R., & Raghavendra, A. S. (2013). Emerging concept for the role of photorespiration as an important part of abiotic stress response. *Plant biology*, 15(4), 713-722.

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

Yano, K., Yamamoto, E., Aya, K., Takeuchi, H., Lo, P. C., Hu, L.,Yoshida, S., Kitano, H., Hirano, K. & Matsuoka, M. (2016). Genome-wide association study using whole-genome sequencing rapidly identifies new genes influencing agronomic traits in rice. *Nature genetics*, 48(8), 927-934.

Yu J., Chen S., Zhao Q., Wang T., Yang C., Diaz C., Sun G., & Dai S. (2011). Physiological and proteomic analysis of salinity tolerance in *Puccinellia tenuiflora. Journal of Proteome Research*, 10, 3852–3870.

Zhan, J., & Meyers, B. C. (2023). Plant small RNAs: their biogenesis, regulatory roles, and functions. *Annual Review of Plant Biology*, *74*, 21-51.

Zhang, C., Xu, G., Huang, R., Chen, C., & Meng, J. (2004). A dominant gdcP-specific marker derived from *Moricandia nitens* used for introducing the C_3-C_4 character from *M. nitens* into Brassica crops. *Plant breeding, 123*(5), 438-443.

Zhang, Y., Massel, K., Godwin, I. D., & Gao, C. (2018). Applications and potential of genome editing in crop improvement. *Genome biology*, *19*(1), 210.

Zhao, H., Zhang, W., Zhang, T., Lin, Y., Hu, Y., Fang, C., & Jiang, J. (2020). Genome-wide MNase hypersensitivity assay unveils distinct classes of open chromatin associated with H3K27me3 and DNA methylation in *Arabidopsis thaliana*. *Genome biology*, 21(1), 1-17.

Zhu, X., Liu, J., Sun, X., Kuang, C., Liu, H., Zhang, L., ... & Hua, W. (2022). Stress-induced higher vein density in the C_3–C_4 intermediate *Moricandia suffruticosa* under Drought and Heat Stress. *Journal of Experimental Botany, 73*(18), 6334-6351.

Zhu, M., He, Q., Lyu, M., Shi, T., Gao, Q., Zhi, H., ... & Gao, Y. (2023). Integrated genomic and transcriptomic analysis reveals genes associated with plant height of foxtail millet. *The Crop Journal*, 11(2), 593-604.

Zhu, X. G., Lyu, M. J., Du, H., Yao, H., Chen, G., Huang, Y., ... & Lu, T. (2024). Integrated genomic analysis revealed a dominant role of transcriptional regulation during the evolution of C_4 photosynthesis in *Flaveria* species.

Zicola, J., Liu, L., Tänzler, P., & Turck, F. (2019). Targeted DNA methylation represses two enhancers of FLOWERING LOCUS T in Arabidopsis thaliana. *Nature plants*, 5(3), 300-307.

Acknowledgments

- Prof. Dr. Benjamin Stich for a life changing opportunity and the years long supervision of this thesis.
- Dr. Suresh Venkata Bonthala and Dr. David Ries for the Bioinformatical guidance.
- Dr. Urte Schüter, Dr. Sebastian Triesch and Prof. Dr. Andreas Weber for the productive and engaging collaboration.
- Stephanie Krey and Agata Stoltmann for the valuable technical assistance at the laboratory.
- My colleagues at the QGGP institute, especially Dr. Marius Weisweiler, Dr. Ruth Freiere Álvarez, Dr. Maria Schmidt and Nadia Baig for the close collaborations.
- To my wife Kati for the inspiration, patience and support throughout the years.
- To my family and friends for always being there.
- To baby Roman for allowing me to work on the dissertation and shining light into my life.