





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

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

<sup>1</sup>Institute of Quantitative Genetics and Genomics of Plants, Faculty of Mathematics and Natural Sciences, Heinrich Heine University, Düsseldorf, Germany

<sup>2</sup>Institute of Plant Biochemistry, Faculty of Mathematics and Natural Sciences, Heinrich Heine University, Düsseldorf, Germany

<sup>3</sup>Cluster of Excellence on Plant Sciences (CEPLAS), Düsseldorf, Germany

<sup>4</sup>Biosystematics Group, Department of Plant Sciences, Wageningen University, Wageningen, The Netherlands

<sup>5</sup>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](mailto: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 C<sub>4</sub> and C<sub>3</sub>-C<sub>4</sub> 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 (C<sub>3</sub> and C<sub>3</sub>-C<sub>4</sub>) 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 C<sub>3</sub>-C<sub>4</sub> 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 C<sub>3</sub>-C<sub>4</sub> intermediate photosynthesis has evolved five times independently. Furthermore, our study provides the first genomic support of the hypothesis that *Diptotaxis 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 C<sub>3</sub>-C<sub>4</sub> intermediate photosynthesis.

## KEYWORDS

Brassicaceae, C<sub>3</sub>-C<sub>4</sub> intermediate photosynthesis, evolution

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2023 The Authors. *Plant, Cell & Environment* published by John Wiley & Sons Ltd.

## 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<sub>2</sub> 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<sub>2</sub> 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<sub>2</sub>. 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<sub>2</sub> and NH<sub>3</sub> (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<sub>2</sub> release in these cells, creating an elevated CO<sub>2</sub> 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](http://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: "ILLUMINA CLIP: 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, atpl-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

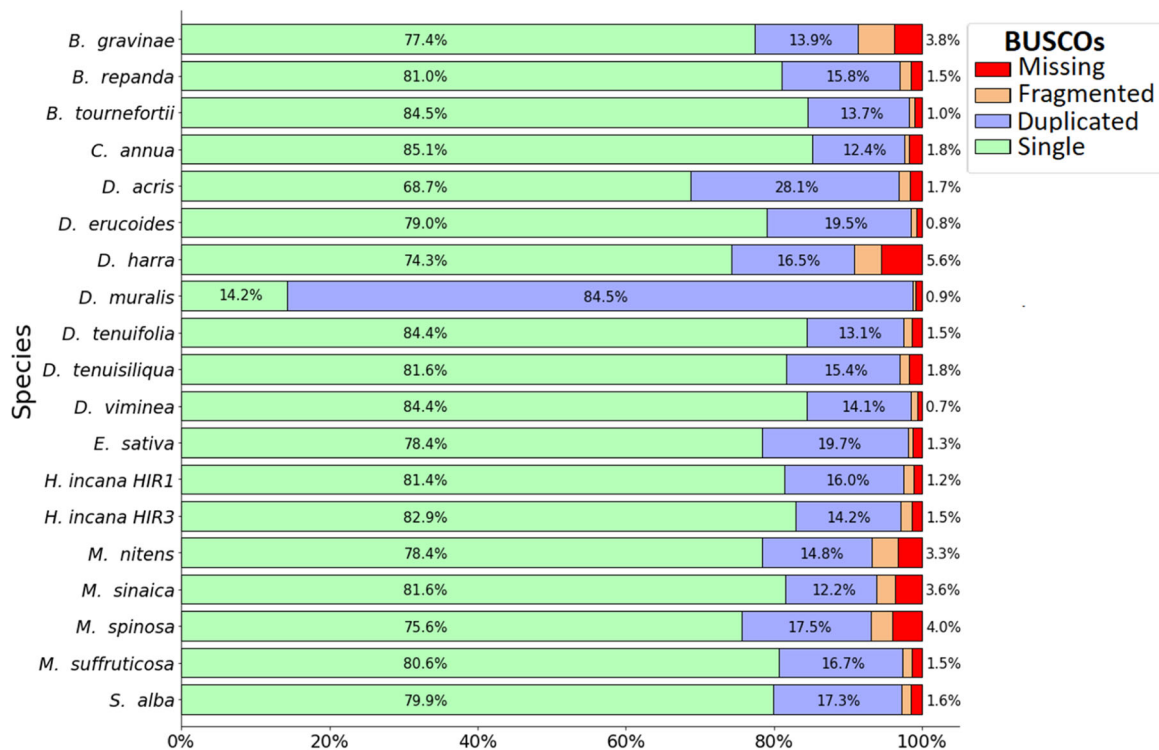
## 3.1 | De novo genome assemblies

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<sub>2</sub> 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
<i>B. gravinae</i>	37.22	C3-C4	49 186	448 675 291	16 594	7247	91%	33 701
<i>B. repanda</i>	55.60	C3	15 286	616 556 673	594 246	231	97%	44 116
<i>B. tournefortii</i>	47.46	C3	933	271 107 445	14 406 319	7	98%	34 193
<i>C. annua</i>	55.33	C3	3696	483 163 416	4 251 916	36	98%	37 608
<i>D. acris</i>	55.57	C3	56 348	884 178 440	93 034	2082	97%	84 702
<i>D. eruroides</i>	30.30	C3-C4	4600	359 627 687	1 993 050	46	98%	44 701
<i>D. harra</i>	53.25	C3	62 651	536 054 563	77 558	1388	91%	55 864
<i>D. muralis</i>	35.04	C3-C4	839	435 802 975	3 038 157	38	99%	84 301
<i>D. tenuifolia</i>	12.41	C3-C4	6128	552 030 198	1 665 732	85	97%	44 007
<i>D. tenuisiliqua</i>	49.38	C3	16 194	510 911 626	328 653	315	97%	40 352
<i>D. viminea</i>	51.08	C3	956	304 993 883	3 734 896	22	98%	38 868
<i>E. sativa</i>	51.82	C3	1371	595 848 844	1 142 695	120	97%	46 211
<i>H. incana HIR1</i>	50.50	C3	6478	371 040 007	885 325	73	97%	43 237
<i>H. incana HIR3</i>	38.37	C3-C4	3768	347 540 957	662 613	120	97%	42 185
<i>M. nitens</i>	21.04	C3-C4	28 171	516 121 665	48 092	3575	93%	41 068
<i>M. sinaica</i>	23.90	C3-C4	44 586	825 206 170	195 836	1041	93%	27 349
<i>M. spinosa</i>	17.80	C3-C4	38 590	443 735 849	35 353	1831	94%	63 573
<i>M. suffruticosa</i>	24.87	C3-C4	9798	516 708 898	991 320	120	97%	41 892
<i>S. alba</i>	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. tournefortii* 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).

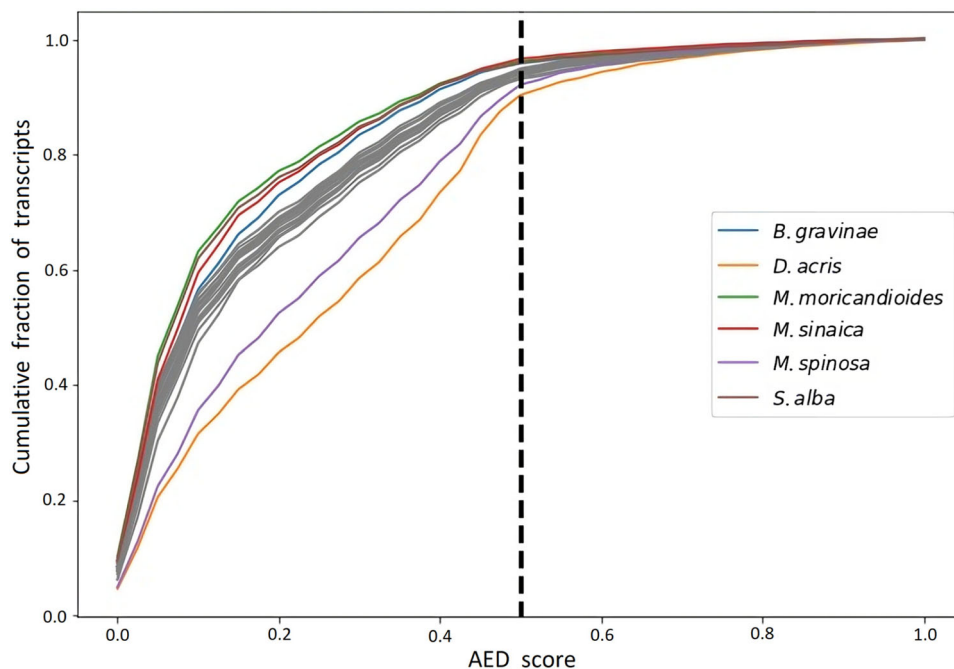
### 3.4 | Repeat annotation

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](https://onlinelibrary.wiley.com/doi/10.1111/pce.14662)]

(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).

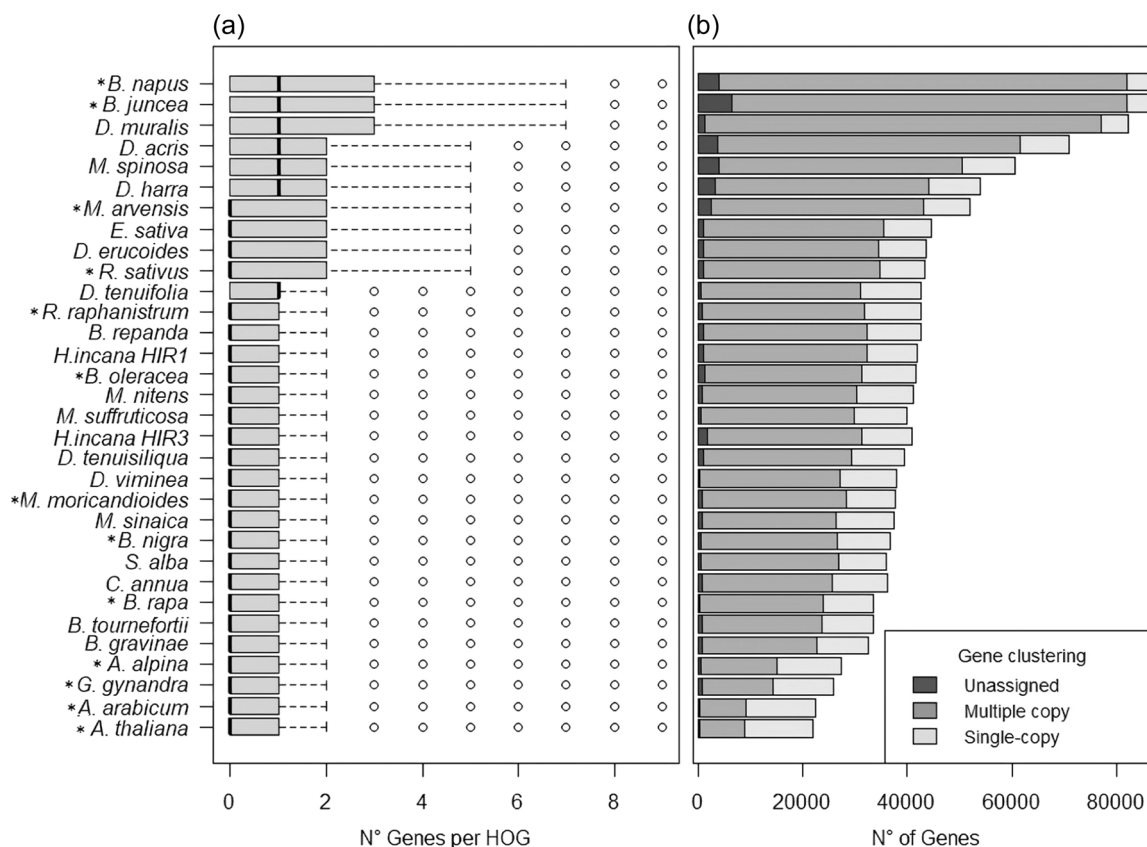
### 3.6 | Orthology

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).

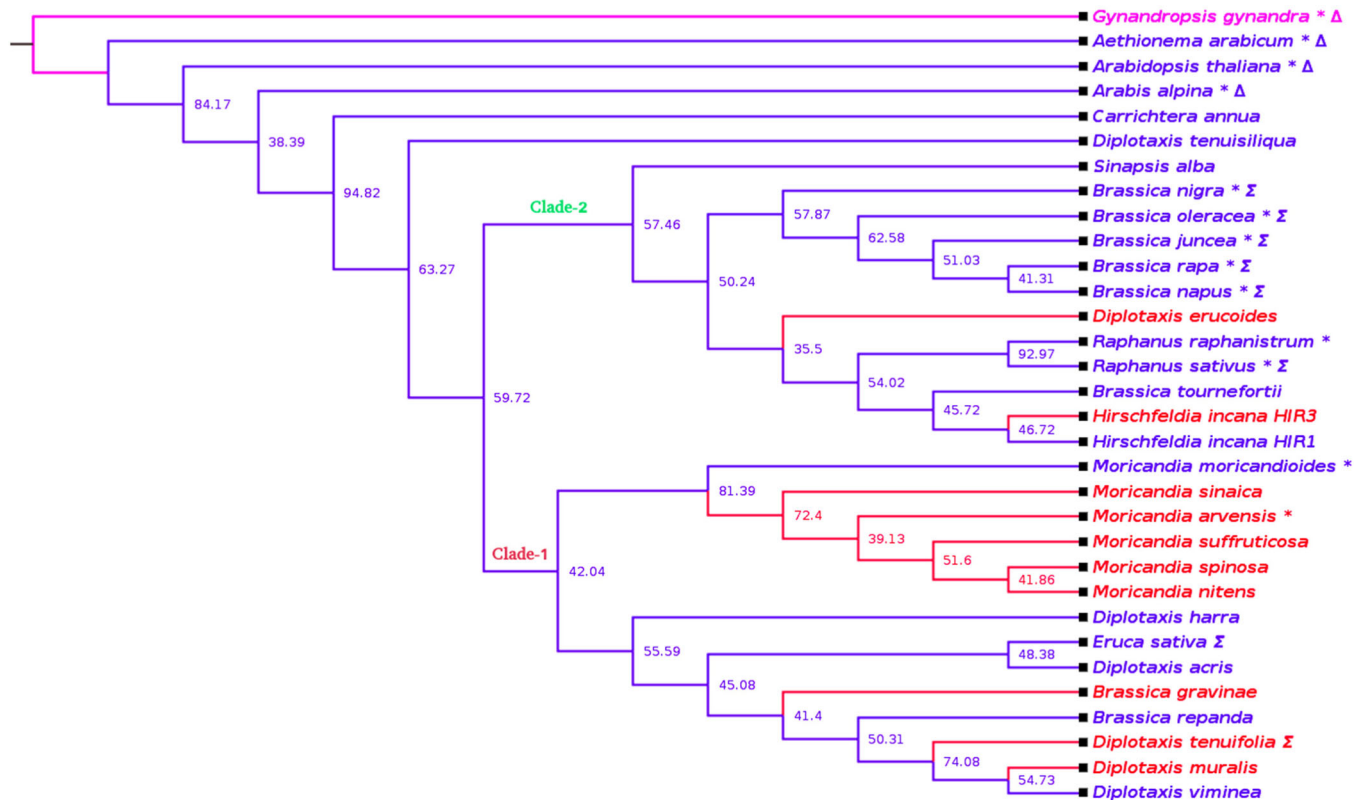
### 3.7 | Genome-wide phylogenetic tree

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<sub>2</sub> 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 (\*).





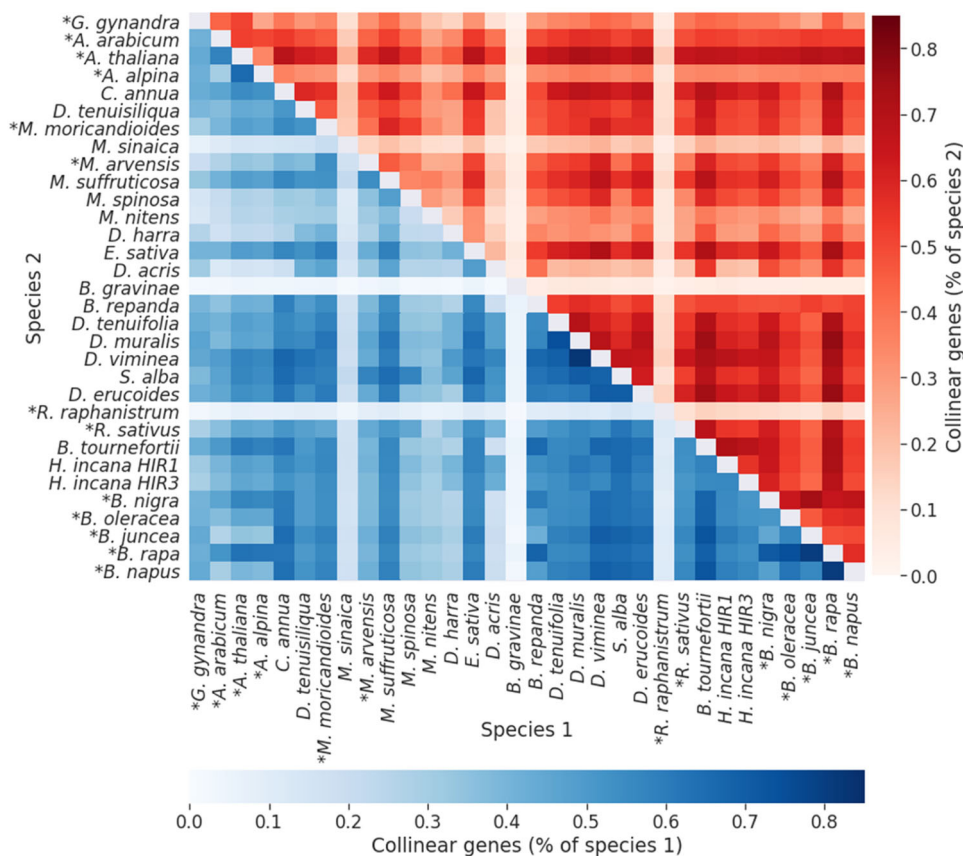
**FIGURE 4** Species tree created using a multi-species coalescent-based approach with *G. gynandra* (C4 photosynthesis) as an outgroup species. Node values are quartet scores created by Astral-Pro, indicating branching support on a 0–100 range, representing the percentage of quartets in the gene trees 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 (Σ), respectively. [Color figure can be viewed at [wileyonlinelibrary.com](https://onlinelibrary.wiley.com/doi/10.1111/pce.14662)]

### 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 Brassicaceae chloroplast genomes had high support values for all nodes (SH-aLRT/bootstrapped >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/bootstrapped = 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/bootstrapped >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](https://onlinelibrary.wiley.com/doi/10.1111/pce.14662)]

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

## 4 | DISCUSSION

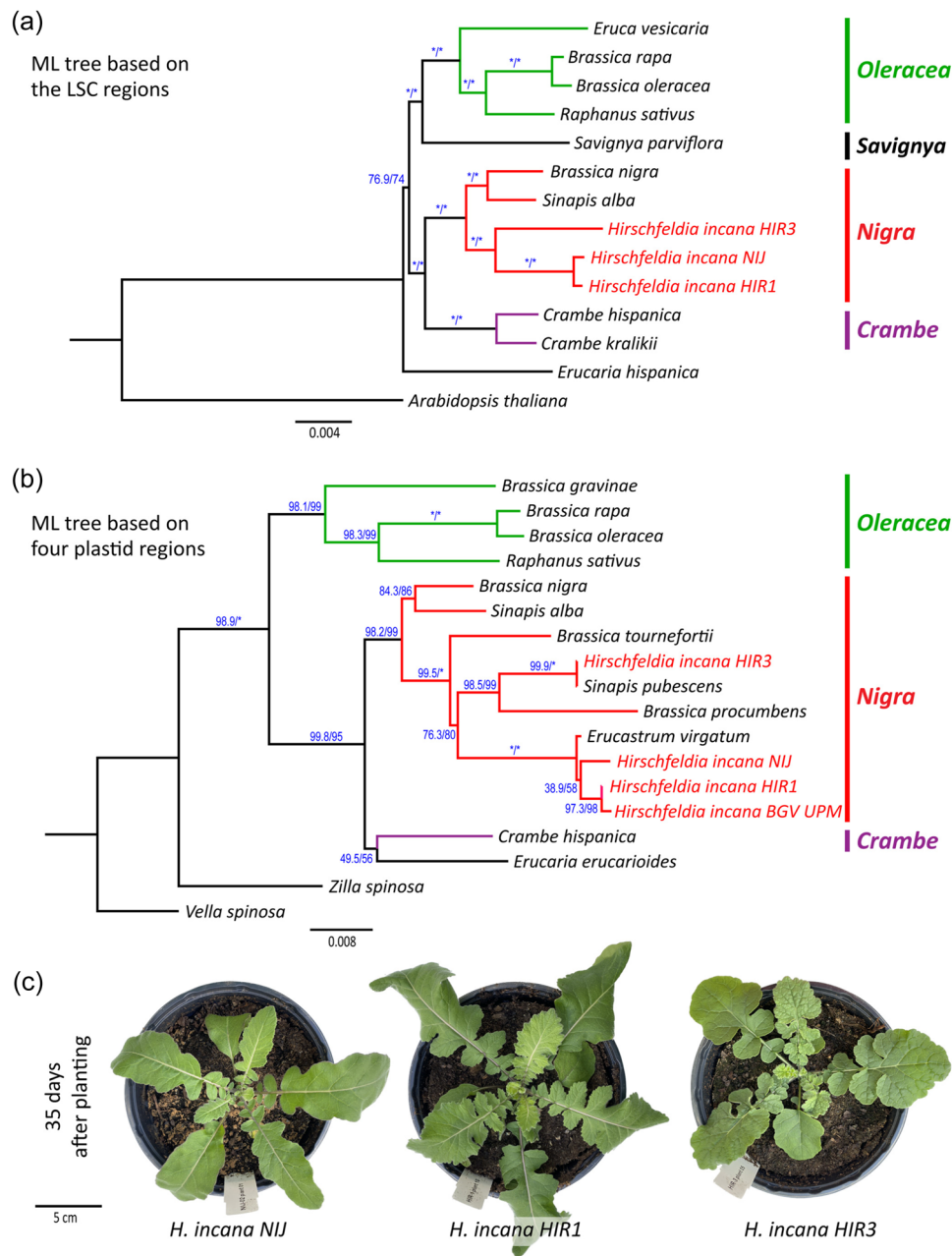
The main aim of this study was to establish resources that enable genomic comparisons and investigations of the evolution of C3-C4 intermediate photosynthesis within the Brassicaceae family and especially the Brassiceae tribe. For such analyses, not only dense sampling of C3-C4 intermediate species but 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](https://onlinelibrary.wiley.com/doi/10.1111/pce.14662)]

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  $\leq$  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 Brassicaceae 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 disregarding 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 autopolyploid. 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 Brassicaceae 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 Brassicaceae 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 Brassicaceae tribe. Since the Brassicaceae 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 Brassicaceae 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 Brassicaceae 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. eruroides* and HIR3, in the same subclade where the commercially important species of the Brassica and Raphanus genera reside (Figure 4). Therefore, *D. eruroides* 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. eruroides* 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<sub>2</sub> 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<sub>2</sub> 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 Gettersleben for providing 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](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](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/tbj.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: Brassicaceae): 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/tbj.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., Kūlahoglu, 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 phylogenetic 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., Wijffjes, 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 blend 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](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. & Jermini, 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<sub>3</sub>–C<sub>4</sub> 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<sub>4</sub> 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 C3–C4 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/pcw009>
- 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<sub>2</sub> 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 C3–C4 intermediate and a C3 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<sub>2</sub> 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 Kranz 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>
- Weiβ, 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>