
Article:
Alenazi, A.S. orcid.org/0000-0003-4105-2539, Pereira, L. orcid.org/0000-0001-5184-8587, Christin, P. orcid.org/0000-0001-6292-8734 et al. (2 more authors) (2024) Identifying genomic regions associated with C4 photosynthetic activity and leaf anatomy in Alloteropsis semialata. New Phytologist, 243 (5). pp. 1698-1710. ISSN 0028-646X

https://doi.org/10.1111/nph.19933

New Phytologist

Research

Identifying genomic regions associated with C_4 photosynthetic activity and leaf anatomy in *Alloteropsis semialata*

Ahmed S. Alenazi[1,2]*, Lara Pereira[2]*, Pascal-Antoine Christin[2], Colin P. Osborne[3] and Luke T. Dunning[2]

[1]Department of Biological Sciences, College of Science, Northern Border University, Arar, 91431, Saudi Arabia; [2]Ecology and Evolutionary Biology, School of Biosciences, University of Sheffield, Western Bank, Sheffield, S10 2TN, UK; [3]Plants, Photosynthesis and Soil, School of Biosciences, University of Sheffield, Western Bank, Sheffield, S10 2TN, UK

Author for correspondence:
Luke T. Dunning
Email: l.dunning@sheffield.ac.uk

Received: *18 March 2024*
Accepted: *13 June 2024*

New Phytologist (2024)
doi: 10.1111/nph.19933

Key words: *Alloteropsis semialata*, bundle sheath, C_4 photosynthesis, genome-wide association study (GWAS), Poaceae.

Summary

- C_4 photosynthesis is a complex trait requiring multiple developmental and metabolic alterations. Despite this complexity, it has independently evolved over 60 times. However, our understanding of the transition to C_4 is complicated by the fact that variation in photosynthetic type is usually segregated between species that diverged a long time ago.
- Here, we perform a genome-wide association study (GWAS) using the grass *Alloteropsis semialata*, the only known species to have C_3, intermediate, and C_4 accessions that recently diverged. We aimed to identify genomic regions associated with the strength of the C_4 cycle (measured using $\delta^{13}C$), and the development of C_4 leaf anatomy.
- Genomic regions correlated with $\delta^{13}C$ include regulators of C_4 decarboxylation enzymes (*RIPK*), nonphotochemical quenching (*SOQ1*), and the development of Kranz anatomy (*SCARECROW-LIKE*). Regions associated with the development of C_4 leaf anatomy in the intermediate individuals contain additional leaf anatomy regulators, including those responsible for vein patterning (*GSL8*) and meristem determinacy (*GIF1*).
- The parallel recruitment of paralogous leaf anatomy regulators between *A. semialata* and other C_4 lineages implies the co-option of these genes is context-dependent, which likely has implications for the engineering of the C_4 trait into C_3 species.

Introduction

Oxygenic photosynthesis originated over 2 billion years ago and is the ultimate source of nearly all energy used by living organisms. Almost 90% of plants fix carbon using the ancestral C_3 cycle, but this process is inefficient in hot environments (Sage & Monson, 1999). This is because the key enzyme responsible for the initial fixation of atmospheric CO_2 (Ribulose-1,5-bisphosphate carboxylase/oxygenase, Rubisco) is less able to discriminate CO_2 from O_2 at higher temperatures, and as a result, energy is lost through photorespiration (Farquhar *et al.*, 1982). To reduce photorespiration, plants have evolved C_4 photosynthesis, wherein atmospheric CO_2 is initially assimilated into a 4-carbon organic acid by phosphoenolpyruvate carboxylase (PEPC) in the mesophyll cells, before shuttling the acid to the neighboring bundle sheath cells where it is decarboxylated and the CO_2 recaptured by Rubisco (Hatch, 1971; Edwards & Ku, 1987). This compartmentalization of Rubisco effectively prevents photorespiration. C_4 photosynthesis is a complex trait that relies on both changes to the leaf anatomy and the coordinated regulation of multiple metabolic enzymes (Hatch, 1987). In order to understand the sequence of events that led to C_4 evolution, comprehensive genomic and phenotypic datasets have been generated in many systems, such as *Flaveria* (Adachi *et al.*, 2023) and *Alloteropsis* (Pereira *et al.*, 2023). These existing data sets can potentially be mined for quantitative genetics approaches to identify novel genetic factors involved in the evolution of C_4 (Simpson *et al.*, 2021).

By comparing species with different photosynthetic types, the core C_4 enzymes, multiple accessory genes, and loci associated with C_4 leaf anatomy (often termed 'Kranz' anatomy) have been identified (Langdale *et al.*, 1987, 1988; Slewinski *et al.*, 2012; Cui *et al.*, 2014). However, decomposing the individual steps during the transition to C_4 is confounded by the fact that variation in photosynthetic type is usually segregated between distinct species that have been independently evolving for millions of years, meaning that they differ in many aspects besides those linked to the photosynthetic pathway (Heyduk *et al.*, 2019). The interspecific segregation of variation in photosynthetic type makes it challenging to apply quantitative genetics methods, such as quantitative trait loci (QTL) mapping and genome-wide association studies (GWAS), since these rely on traits varying within a species, or the ability to hybridize species with divergent phenotypes. GWAS has been used to investigate the variation of C_4 traits within C_4 species, such as photosynthetic performance during chilling in maize (Strigens *et al.*, 2013), and to identify genes

*These authors contributed equally to this work.

associated with stomatal conductance and water use efficiency in sorghum (Ferguson *et al.*, 2021; Pignon *et al.*, 2021). However, to date there has been no QTL region identified for differences in C_4 carbon fixation or Kranz anatomy (Simpson *et al.*, 2021).

In grasses, the proportion of carbon that is fixed through the C_4 cycle can be measured using the stable carbon isotope ratio ($\delta^{13}C$) (O'Leary, 1981; Farquhar *et al.*, 1989). Both ^{12}C and ^{13}C occur naturally in the atmosphere, and in C_3 plants, Rubisco preferentially fixes ^{12}C during photosynthesis (O'Leary, 1981). Conversely, in C_4 plants, carbon is initially fixed by CA and PEPC, and this coupled enzyme system discriminates less than Rubisco between the two isotopes (O'Leary, 1981). The rate of CO_2 release in the bundle sheath is coordinated with the rate of CO_2 fixation by Rubisco, which reduces the fractionation effect of this enzyme. $\delta^{13}C$ is therefore commonly used as a proxy for photosynthetic type and the relative strength of the C_4 cycle (Bender, 1968; Smith & Epstein, 1971; Smith & Brown, 1973; Von Caemmerer, 1992; Cerros-Tlatilpa & Columbus, 2009; Gowik *et al.*, 2011; Lundgren *et al.*, 2015; Stata *et al.*, 2019; Olofsson *et al.*, 2021). While there is intraspecific variation in $\delta^{13}C$ for C_4 species such as maize and *Gynandropsis* (Voznesenskaya *et al.*, 2007), we do not know whether this variation arises from differences in anatomy or biochemistry (Simpson *et al.*, 2021). In addition, some of the observed variation in $\delta^{13}C$ could also be due to environmental effects on water use efficiency (Farquhar & Richards, 1984), particularly if the phenotypic data comes from individuals sampled in the field. However, differences in the $\delta^{13}C$ between accessions of some species are maintained in a common environment (Lundgren *et al.*, 2016), indicating that the $\delta^{13}C$ ratio likely has a genetic component. Intraspecific, heritable variation in $\delta^{13}C$ offers an excellent opportunity for using quantitative genetic approaches to discover C_4 QTLs.

The grass *Alloteropsis semialata* has long been used as a model to study C_4 evolution, since it is the only species known to have C_3, C_4, and intermediate genotypes that diverged relatively recently and can be crossed, allowing gene flow among them (reviewed by Pereira *et al.*, 2023). The common ancestor of this species is thought to be an intermediate with some chloroplasts in its bundle sheath and performing a very weak C_4 cycle, with the C_3 being a reversal from this intermediate state as that lineage colonized cooler environments in southern Africa (Dunning *et al.*, 2017). The intermediate populations are found in the grassy ground layer of the Central Zambezian miombo forests that we refer to as 'C_3+C_4' because they perform a weak C_4 cycle in addition to directly fixing CO_2 through the C_3 cycle (Lundgren *et al.*, 2016; Dunning *et al.*, 2017). Comparative studies have shown that the transition to a purely C_4 physiology in *A. semialata* is caused by the overexpression of relatively few core C_4 enzymes (Dunning *et al.*, 2019a) and the acquisition of C_4-like morphological traits, notably the presence of minor veins (Lundgren *et al.*, 2019). The $\delta^{13}C$ of the C_3+C_4 plants ranges from values characteristic of a weak (or absent) C_4 cycle to values that show that the C_4 cycle accounts for more than half of the carbon acquisition (Von Caemmerer, 1992; Lundgren *et al.*, 2015; Stata *et al.*, 2019; Olofsson *et al.*, 2021). Furthermore, the strengthening of the C_4 cycle in the C_3+C_4 intermediates (measured using $\delta^{13}C$) is significantly associated with alterations in a number of leaf anatomical traits related to the preponderance of inner bundle sheath (IBS) tissue, the cellular location of the C_4 cycle in this species (Alenazi *et al.*, 2023), including the distance between consecutive bundle sheaths, the width of IBS cells, and the proportion of bundle sheath tissue in the leaf (Alenazi *et al.*, 2023).

Alloteropsis semialata therefore represents an ideal system to identify the genes correlated with the strengthening of the C_4 cycle. Here, we first conducted a global analysis to identify candidate genes associated with the strength of the C_4 cycle ($\delta^{13}C$) using genomic data from 420 individuals representing C_3, C_3+C_4, and C_4 phenotypes. We then focused specifically on the C_3+C_4 intermediates, to identify candidate genes associated with the relative expansion of bundle sheath tissue during the transition from a weak to a strong C_4 cycle. The high level of interspecific variation in *A. semialata* permits a fine-scale understanding of the genetic basis of C_4 evolution, including the intermediate steps involved in the assembly of this complex trait. This is crucially important to identify the initial changes required for the emergence of this trait, something that may ultimately have applications in the engineering of C_4 photosynthesis in C_3 crops such as rice.

Materials and Methods

Genome data, $\delta^{13}C$ values, and population genetic analyses

For the genomic analyses, we compiled previously published double digest restriction-site associated DNA sequencing (ddRADSeq) data sets for *Alloteropsis semialata* (R. Br.) Hitchc. individuals that also had known $\delta^{13}C$ values from field-collected leaves measured using mass spectrometry (Lundgren *et al.*, 2015, 2016; Bianconi *et al.*, 2020; Olofsson *et al.*, 2021; Alenazi *et al.*, 2023). Depending on the source of the $\delta^{13}C$ values, these were either single measures (Lundgren *et al.*, 2015, 2016; Bianconi *et al.*, 2020), replicated if the $\delta^{13}C$ values did not match other individuals of the population and genomic group (Olofsson *et al.*, 2021), or medians of triplicate technical replicates if sufficient material was available (Alenazi *et al.*, 2023). In total, the data set comprised 420 individuals collected from 87 populations across Africa and Asia (Supporting Information Table S1), representing the full range of photosynthetic types found in *A. semialata* (45 × C_3; 132 × C_3+C_4; 243 × C_4).

The ddRADseq data were downloaded from NCBI Sequence Read Archive and cleaned using Trimmomatic v.0.38 (Bolger *et al.*, 2014) to remove adapter contamination (ILLUMINACLIP option in palindrome mode) and low-quality bases ($Q < 3$ from both 5′ and 3′ ends; $Q < 15$ for all bases in four-base sliding window). The cleaned ddRADseq data were then mapped to a chromosomal scale *A. semialata* reference genome for a C_4 Australian individual (Dunning *et al.*, 2019b) using Bowtie2 v.2.2.3 with default parameters (Langmead & Salzberg, 2012). We called single-nucleotide polymorphisms (SNPs) from these alignments

using the GATK v.3.8 (McKenna *et al.*, 2010) pipeline with default parameters. We generated individual variant files (gVCF) with HAPLOTYPECALLER and then combined them into a single multi-sample VCF file with Genotype GVCFs. Biallelic SNPs were extracted from this file using SELECTVARIANTS, and high-quality SNPs retained using VARIANTFILTRATION (MQ > 40, QD > 5, FS < 60, MQRankSum > −12.5 ReadPosRankSum > −8). Finally, we used VCFTOOLS to filter the remaining SNPs to remove those with > 30% missing data and/or a minor allele frequency < 0.05 (Danecek *et al.*, 2011).

The evolutionary relationship among individuals was inferred using a maximum likelihood phylogenetic tree. We used VCF2PHYLIP v.2.8 (Ortiz, 2019) to generate a nucleotide alignment from the filtered VCF file. To reduce the effect of linked SNPs on phylogenetic reconstruction, we thinned the data set so that SNPs were at least 1 kb apart (starting from the first SNP on each chromosome). The phylogenetic tree was inferred using RAXML v.8.2.12 (Stamatakis, 2014) with the GTRCAT model and 100 bootstrap replicates (Dataset S1). Finally, to verify previous phylogenetic groupings (Alenazi *et al.*, 2023), we determined the population structure of the C_3+C_4 accessions using ADMIXTURE v.1.3.0 (Alexander *et al.*, 2009). We ran the analysis with multiple values of k (range: 2–7), with 10 replicate runs for each value. The optimal k was inferred using Admixture's cross-validation error method. We also used PLINK-v.1.9 to perform a principal component analysis to quantify population structure and to generate a pairwise kinship matrix (Purcell *et al.*, 2007).

Leaf anatomical traits of C_3+C_4 *A. semialata*

Leaf anatomy data for all 132 C_3+C_4 individuals were either extracted from a previous study (n = 100; Alenazi *et al.*, 2023) or generated here using the same method from field-preserved samples (n = 32; Table S2). The measurements themselves were taken from leaf cross-sections that were prepared from silica-dried leaf material following the method described by Alenazi *et al.* (2023). The slide images were captured using a mounted camera on an Olympus BX51 microscope (Olympus, Hamburg, Germany), and images from the same leaf were stitched together with Hugin's software (Hugin Development Team, 2015). All measurements of leaf anatomical characteristics were made using IMAGEJ v.1.53f (Schneider *et al.*, 2012), avoiding the midrib and leaf margins. For each individual, the anatomical measurements are based on the mean of at least five technical replicates, each measured between independent pairs of secondary veins in the same cross-section using.

We recorded the total cross-sectional areas between secondary veins (i.e. veins accompanied by extraxylary fibers and epidermal thinning) for mesophyll (including airspaces; MS) and IBS tissues (Fig. 1). We used these values to then calculate the inner bundle sheath fraction (IBSF = IBS/[MS + IBS]), which is the portion of the photosynthetic part of the leaf that can be responsible for refixing carbon obtained through the C_4 cycle. Finally, we also measured the bundle sheath distance (BSD) and the inner bundle sheath width (IBSW) using the mean widths of equatorial cells.

Estimating trait heritability

To estimate the proportion of phenotypic variation explained by underlying genetic differences, we calculated the heritability of $\delta^{13}C$ (complete and restricted C_3+C_4 datasets) and the leaf anatomical traits (C_3+C_4 dataset) using Genome-wide Complex Trait Analysis (GCTA) v.1.94.1 (Yang *et al.*, 2011). A genetic relationship matrix was inferred from the previously generated SNP calls and combined with the phenotype values in GCTA. Heritability was then estimated for each trait using the restricted maximum likelihood method.

Genome-wide association study of photosynthetic traits

We performed a GWAS for the strength of the C_4 cycle measured using $\delta^{13}C$, and leaf anatomy traits previously correlated with the strength of the C_4 cycle (IBSF, BSD, and IBSW; Fig. 1; Alenazi *et al.*, 2023), with the objective of ultimately proposing some candidate genes underpinning the phenotype. We used the variation in photosynthetic type which exists across *A. semialata* as a whole, before focusing on anatomical variation in the C_3+C_4 individuals that have been associated with the strength of the C_4 cycle (Alenazi *et al.*, 2023). We defined our associated regions of the genome as the linkage block containing a significant SNP from the GWAS. We then identified the gene models located within the correlated region as candidate genes and assessed their functional relevance, gene expression pattern, and selective forces they have been evolving under.

The GWAS itself was performed using the RMVP package (Yin *et al.*, 2021) in RSTUDIO v.4.3, with the MVP.Data function and default parameters used for single-locus GWAS analysis for each phenotypic trait with the fixed and random model circulating probability unification (FarmCPU) approach (Yin *et al.*, 2021). Population structure and genetic relatedness can confound a GWAS and result in false associations (Chen *et al.*, 2016). We therefore included the previously generated pairwise kinship matrix so that the relationships among individuals could be accounted for. The phenotypic data for each trait were normalized (if required), and a Bonferroni corrected SNP significance threshold of $P \leq 0.05$ was used.

Linkage disequilibrium

Linkage blocks are regions of the genome that are likely to be co-inherited, and the association of the significant SNPs identified from the GWAS could be caused by any gene within this region. To determine the linkage block encapsulating each SNP, we used HAPLOVIEW v.4.1 (Barrett *et al.*, 2005). The input map and binary files were processed using PLINK v.1.9 (Purcell *et al.*, 2007), and we used a solid spine of linkage disequilibrium (LD) with default parameters to infer linkage block size (Kim *et al.*, 2018). This approach requires the first and last SNPs in a block to be in strong LD with all intermediate markers (normalized deviation (D') ≥ 0.8), but the intermediate markers do not necessarily need to be in LD with each other. Identifying linkage blocks is heavily impacted by the distribution of SNPs across the genome,

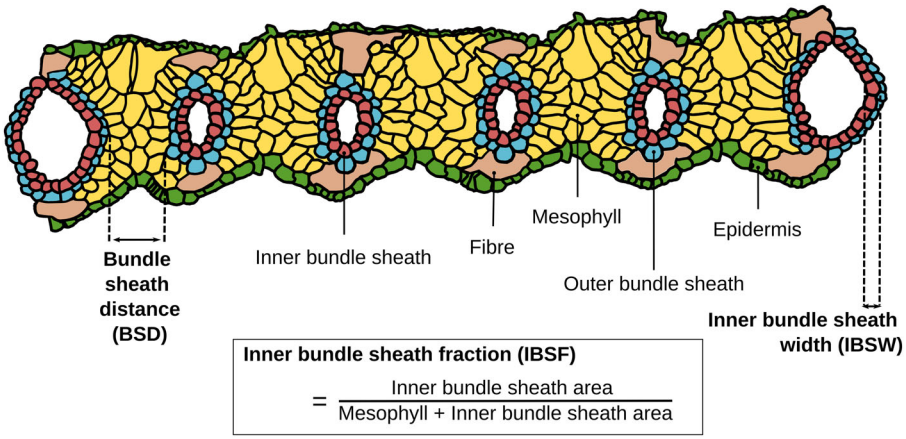

Fig. 1 Cross-section of a typical C_4 *Alloteropsis semialata* leaf. The schematic is traced from a cross-section of individual JKO23-03_16, with tissue types labelled. The anatomical measurements used for the genome-wide association study (GWAS) are indicated in bold.

something that is accentuated by reduced sequencing methods such as ddRADSeq. We therefore used the genome-wide mean linkage block size if the analysis failed to place a significant SNP in a block of its own (Fig. S1). To do this, we positioned the significant SNP at the center of the artificial linkage block and if necessary truncated it to avoid incorporating unlinked SNPs up and/or downstream from this marker.

Identification of candidate genes

The linkage blocks associated with the phenotype of interest contain the causal gene(s) in addition to those that happen to be in close physical linkage (hitchhiking). To try and identify plausible candidate genes in each region, we compared their functional annotations, expression patterns, and the selective pressures they are evolving under.

ORTHOFINDER v.2.5.4 (Emms & Kelly, 2015) was used to identify orthologous genes to the loci in the associated regions. To do this, we combined the *A. semialata* protein sequences with nine other plant species (*Arabidopsis thaliana*, *Brachypodium distachyon*, *Hordeum vulgare*, *Oryza sativa*, *Physcomitrium patens*, *Solanum lycopersicum*, *Triticum aestivum*, and *Zea mays*) downloaded from Phytozome v.13 (Goodstein *et al.*, 2012). Orthogroup phylogenies are presented in Dataset S2. We then used publicly available databases (e.g. TAIR (Berardini *et al.*, 2015), RAP-DB (Sakai *et al.*, 2013), and maizeGDB (Monaco *et al.*, 2013)) and literature searches to extrapolate the functions of each orthogroup containing a gene from a correlated linkage block identified from the GWAS.

Gene expression data for the candidate genes was extracted from a Dunning *et al.* (2019a). The gene expression data come from mature leaf tissue grown under controlled conditions (60% relative humidity, day/night temperatures of 25/20°C), sampled in the middle of the photoperiod (Dunning *et al.*, 2019a). To test for differential expression between the photosynthetic types, we used two-tailed *t*-tests, with *P*-values Bonferroni corrected to account for multiple testing.

Finally, we used whole-genome resequencing data (Bianconi *et al.*, 2020) for 45 *A. semialata* individuals to determine whether the genes in the GWAS regions were evolving under positive selection. In short, the datasets were downloaded from NCBI sequence read archive and mapped to the reference genome using BOWTIE2, and consensus sequences generated using previously developed methods (Olofsson *et al.*, 2016; Dunning *et al.*, 2022), and a maximum likelihood phylogeny tree for each gene was inferred using RAXML (Stamatakis, 2014) with 100 bootstrap. We then inferred the selective pressure each gene was evolving under by running the M0 model in CODEML v.4.9 h.

Results

Population structure

The broadscale phylogenetic (Fig. 2a) and population genetic (Fig. 2b) analyses recovered those previously inferred by earlier studies, with the different photosynthetic types (C_3, C_3+C_4, and C_4) belonging to separate clades (Olofsson *et al.*, 2016, 2021; Bianconi *et al.*, 2020). Within the C_3+C_4 intermediates, individuals are separated into five populations geographically spread across the Central Zambezian miombo woodlands (Fig. 2). This reconfirms the phylogenetic groupings previously demarcated (Alenazi *et al.*, 2023), although the earliest diverging sixth lineage is absent in this study because it is only represented by a single herbarium individual from the Democratic Republic of the Congo and lacks ddRADSeq data. The population structure analysis (Fig. 2c) largely concurs with the phylogenetic groupings, although it indicates gene flow between populations. The distribution of the C_3+C_4 groups has a pattern largely matching a scenario of isolation by distance along an east–west axis through Zambia and Tanzania (Fig. 2d). *Q*–*Q* plots show no sign of *P*-value inflation for the subsequent GWAS results, indicating that population structure was sufficiently corrected for using the pairwise kinship matrix (Fig. S2).

Identifying regions of the genome correlated with the strength of the C_4 cycle

We used the $\delta^{13}C$ values as a proxy for the strength of the C_4 cycle for all 420 *A. semialata* individuals used in this study. As expected, the $\delta^{13}C$ values supported the demarcation of the main

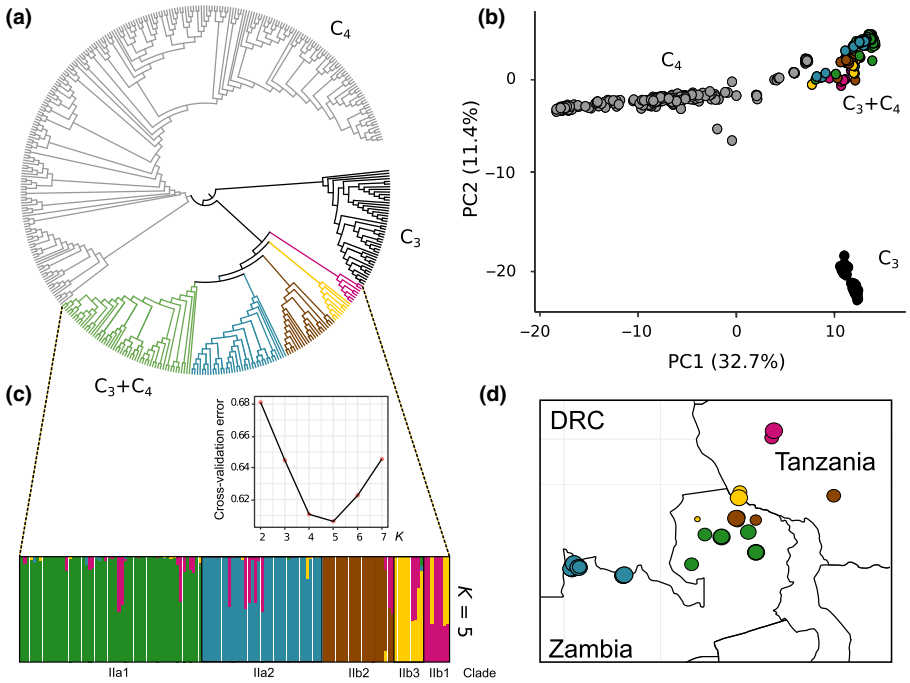

Fig. 2 Population genomics of *Alloteropsis semialata*. (a) A cladogram of the maximum likelihood phylogenetic tree, with individual clades recovered within the C_3+C_4 lineage colored (same colors used in all panels). (b) A principal component analysis of the genotypes, showing the first two axes. (c) Admixture results for the C_3+C_4 *A. semialata* individuals for $K = 5$, the optimal number of population clusters based on the cross-validation error. (d) Location of the C_3+C_4 populations used in this study, with the size of the point proportional to the number of individuals sampled (range: 1–20 individuals per population).

nuclear clades into the C_3, C_3+C_4, and C_4 phenotypes (Fig. 3a). For C_3 and C_4 individuals, we found $\delta^{13}C$ average values of −26.67‰ and −12.63‰ with little dispersion within each group, whereas for C_3+C_4 individuals, we found substantial variation ranging from −28.35‰ to −18.47‰ with an average of −23.87‰. The heritability estimate, which represents the proportion of phenotypic variation due to genetic variation in the population, was high for $\delta^{13}C$ when considering all photosynthetic types ($h^2 = 0.75$; SE = 0.06; $n = 420$), and threefold lower when just considering the C_3+C_4 intermediates ($h^2 = 0.25$; SE = 0.00; $n = 132$). This reduced heritability estimate could be due to multiple factors, including reduced power (132 C_3+C_4 individuals vs 420 full dataset), limited variation (C_3+C_4 range = 9.88‰; full data set range = 19.4‰), or increased residual variation within the C_3+C_4 compared with other photosynthetic types (C_3+C_4 SD = 2.1‰, C_3 SD = 1.08‰, C_4 SD = 1.0‰; Fig. 3a).

We conducted a combined GWAS using all individuals (Fig. 3b), as well as various partitions by photosynthetic type (Fig. S3). When considering all individuals, the GWAS identified three significant SNPs on chromosome 9, which all corresponded to relatively narrow regions based on the LD (Fig. 3b). The region with the highest association with $\delta^{13}C$ (LB-01) is a 121-kb region at 32.2 Mb (Tables 1, S3; Fig. 3b). The same region was also significant when repeating the GWAS within the C_3+C_4, and when combining the C_3+C_4 with either the C_3 or C_4 individuals (Table S3; Fig. S2), but not when excluding the C_3+C_4 individuals. These results imply that the underlying causative gene segregates only within the C_3+C_4 group. There were six predicted protein-coding genes in the LB-01 region, and all were expressed in the leaf tissue of at least one *A. semialata* individual (Table S4). One of these genes (*Shewanella-like protein phosphatase 1*, *SLP1* (*ASEM_AUS1_34305*)) was significantly more highly expressed in the C_3 than in the other photosynthetic types (C_3 vs C_4 Bonferroni-adjusted *P*-value = 0.073; C_3 vs C_3+C_4 Bonferroni-adjusted *P*-value = 0.015; Fig. 3c; Table S4), although there is no consistent differential expression between photosynthetic types when individual populations are compared separately (Dunning *et al.*, 2019a). None of the six genes were found to be strictly evolving under positive selection with a dN/dS ratio (ω) > 1 (Table S4), although those with the highest values may be seeing a relaxation of purifying selection (e.g. ω = 0.83 for *ASEM_AUS1_34303*). The annotated genes in the LB-01 region have a variety of functions (Table S4), including loci associated with the regulation of the Calvin cycle (*SLP1* (*ASEM_AUS1_34302*) (Kutuzov & Andreeva, 2012; Johnson *et al.*, 2020)) and the activation of NADP-malic enzyme 2 (NADP-ME2), a C_4 decarboxylation enzyme (*RPM1-Induced Protein Kinase*, *RIPK* (*ASEM_AUS1_34305*) (Wu *et al.*, 2022)).

The two other regions identified in the $\delta^{13}C$ GWAS using all individuals (LB-02 and LB-03; Fig. 3b) were not significant when partitioning the data by photosynthetic type (Table S3). Both these regions are delimited by LD blocks narrow in size and that contain one annotated gene each. The candidate gene in LB-02 (*ASEM_AUS1_29467*) was not expressed at all in any *A. semialata* mature leaves, while the one in LB-03 (*ASEM_AUS1_14480*) was expressed in all individuals, but was not differentially expressed between photosynthetic types. In addition, both genes do not seem to have been under positive selection (Table S4). One of these genes (*ASEM_AUS1_29467*) encodes a SCARECROW-LIKE protein 9 (SCL9) protein belonging to the GRAS gene family, a group of transcription factors shown to play a key role in C_4 leaf anatomy and photosynthetic development in maize (Slewinski *et al.*, 2012; Hughes &

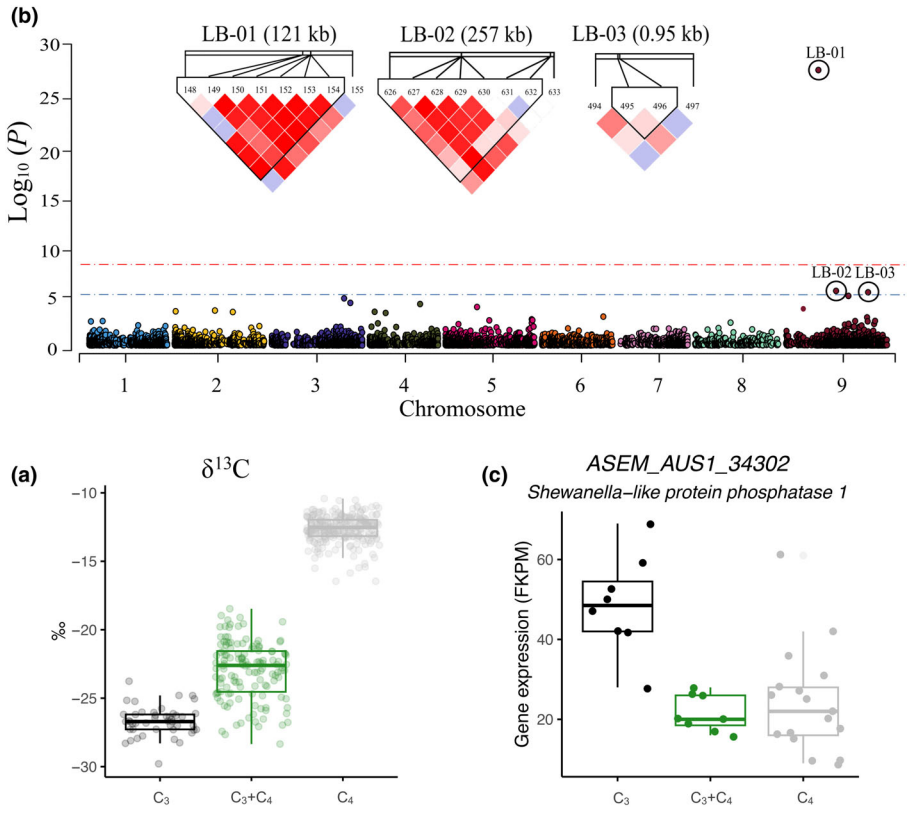

Fig. 3 Genetic variation associated with the strength of the C_4 cycle in *Alloteropsis semialata*. (a) The stable carbon isotope ratio ($\delta^{13}C$) was used to infer the strength of the C_4 cycle, with values measured for each of the photosynthetic types shown. The boxes show the median value and the interquartile range, and the whiskers represent 1.5× the interquartile range. (b) Manhattan plot showing the results of a genome-wide association study (GWAS) for $\delta^{13}C$ using all samples. The blue and red dotted lines indicate Bonferroni corrected *P*-values of 0.05 and 0.001, respectively. Significant single-nucleotide polymorphisms (SNPs) are labelled with a block ID. Heatmap of pairwise linkage disequilibrium (LD) between markers surrounding each significant SNP, ranging from white indicating low LD (LOD < 2 and D′ < 1) to bright red indicating strong LD (LOD ≥2 and D′ = 1). (c) Boxplot of gene expression for the candidate gene that is significantly differentially expressed. The boxes show the median value and the interquartile range, and the whiskers represent 1.5× the interquartile range.

Table 1 Significantly correlated regions of the genome identified in the genome-wide association studies (GWAS).

Phenotype	Chromosome	SNP position	LD block (kb)	$-\log_{10} P$	Bonferroni-adjusted *P*-value	No. of genes
$\delta^{13}C$	9	32 191 256	121	29.18	5.33E-26	6
	9	49 592 498	<1[b]	5.73	1.51E-02	1
	9	81 051 792	63	5.58	2.12E-02	1
IBSF	9	58 663 539	362	6.88	3.58E-04	33
	2	634 463	6[b]	4.83	4.95E-02	2
	5	23 291 361	59[a]	5.89	3.53E-03	4
	4	63 231 217	13	6.71	5.34E-04	2
	8	23 357 684	275	16.47	9.25E-14	24
BSD	9	10 612 628	11	5.10	2.17E-02	1
	9	48 749 591	<1[b]	5.59	7.02E-03	0

Detailed location of the significant single-nucleotide polymorphism (SNP) identified by GWAS and the linkage disequilibrium (LD) block where they are contained within the *Alloteropsis semialata* genome. The phenotypes used to perform the GWAS are carbon isotope ratio ($\delta^{13}C$), internal bundle sheath fraction (IBSF), and bundle sheath distance (BSD).
[a]This SNP was not located in a linkage block in our analyses; we therefore defined the region using the genome-wide median block size.
[b]This SNP was not located in a linkage block in our analyses; we therefore defined the region using the genome-wide median block size that was truncated if there was a closely located unlinked SNP up- or downstream.

Langdale, 2020). The lack of expression (or differential expression) of these candidate genes in transcriptomes generated from mature leaf tissues is likely explained by the involvement of these genes, such as *SCL9*, in leaf development. The other gene encodes a protein associated with the suppression of nonphotochemical quenching and maintaining the efficiency of light harvesting (*suppressor of quenching 1*, *SOQ1* (*ASEM_AUS1_14480*) (Brooks *et al.*, 2013; Duan *et al.*, 2023)).

The $\delta^{13}C$ GWAS analyses were repeated with a subset of photosynthetic types (Fig. S3; Table S4), and these also identified potentially interesting candidate genes, particularly those related to leaf vein patterning. WIP C2H2 zinc finger protein (*WIP2* (*ASEM_AUS1_03361*); LB-20; C_4 & C_3+C_4 individuals) is paralogous to the WIP6 transcription factor TOO MANY LATERALS that specifies vein rank in maize and rice (Vlad *et al.*, 2024). Defectively Organized Tributaries 4 (*DOT4*

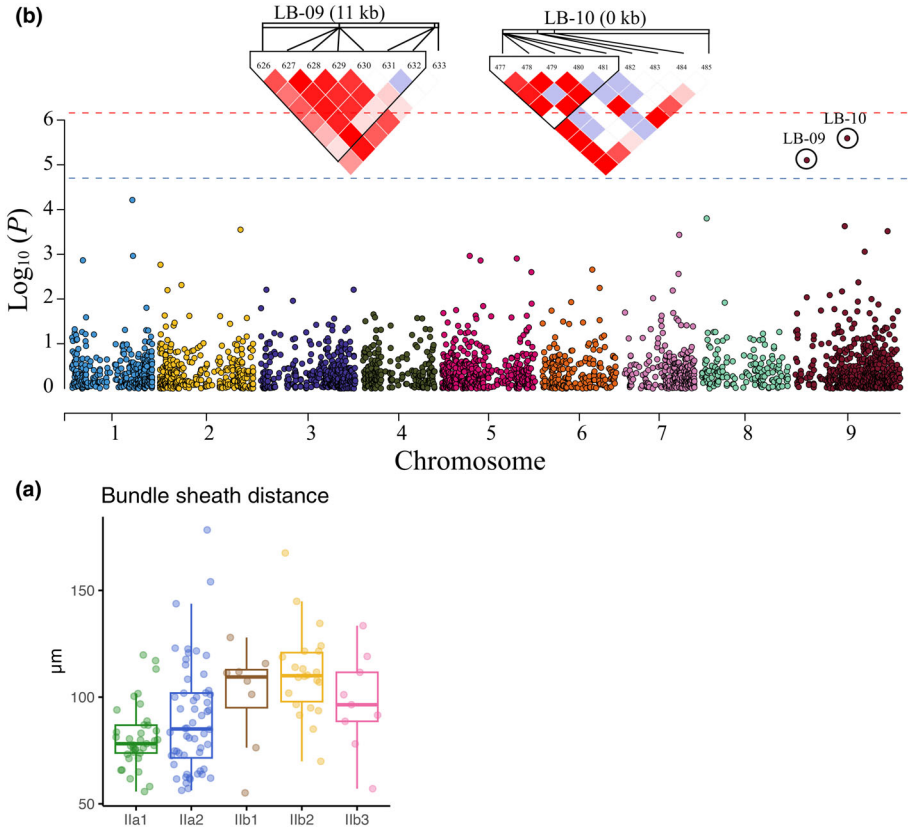

Fig. 4 Genetic variation associated with bundle sheath distance (BSD) in the C_3+C_4 *Alloteropsis semialata*. (a) The boxplot shows the BSD variation for each of the C_3+C_4 subclades. The box indicates the median value and the interquartile range, and the whiskers represent 1.5× the interquartile range. (b) A Manhattan plot showing the results of a genome-wide association study (GWAS) for BSD. The blue and red dotted lines indicate Bonferroni corrected *P*-values of 0.05 and 0.001, respectively. Significant single-nucleotide polymorphisms (SNPs) are labelled with a block ID. Heatmap of pairwise linkage disequilibrium (LD) between markers surrounding each significant SNP, ranging from white indicating low LD (LOD < 2 and D′ < 1) to bright red indicating strong LD (LOD ≥ 2 and D′ = 1).

(*ASEM_AUS1_05127*); LB-23; C_3+C_4 individuals) is orthologous to a vein patterning gene in *Arabidopsis thaliana* (Petricka *et al.*, 2008).

Identifying regions of the genome associated with C_4 leaf anatomy in the C_3+C_4 intermediates

We studied the genetic basis of three leaf anatomical traits previously associated with the strength of the C_4 cycle ($\delta^{13}C$) using the 132 C_3+C_4 individuals (Fig. 1; Alenazi *et al.*, 2023). The heritability estimates for the three leaf anatomical traits in the C_3+C_4 intermediates ranged from roughly equivalent to the value for $\delta^{13}C$ value to much lower (IBSF $h^2 = 0.22$ (SE = 0.04); BSD $h^2 = 0.12$ (SE = 0.06); IBSW $h^2 = 0.06$ (SE = 0.06); $n = 132$). No significantly correlated genomic region was detected for IBSW (Fig. S4), the trait with the lowest heritability. However, we did detect SNPs significantly associated with BSD and IBSF.

Bundle sheath distance The distance between consecutive bundle sheaths plays a significant role in determining the rate and efficiency of photosynthesis in plants, with smaller distances being significantly correlated with higher $\delta^{13}C$ (more C_4-like) values (Alenazi *et al.*, 2023). The C_3+C_4 intermediate individuals showed a range of BSDs from 55.14 to 178.36 µm, with variation between subclades (Fig. 4a). The GWAS identified two significant regions associated with BSD, both on chromosome 9 (Tables 1, S3; Fig. 4). Only one annotated gene was identified in the correlated genomic regions associated with BSD, the function of which is associated with leaf development (*Glucan Synthase-Like 8*, *GSL8* (*ASEM_AUS1_16831*); Table S4 (Linh & Scarpella, 2022)).

Inner bundle sheath fraction Inner bundle sheath fraction represents the portion of the leaf that can be used for C_4 photosynthesis (Fig. 1). A higher IBSF in the C_3+C_4 *A. semialata* has been significantly correlated with a higher $\delta^{13}C$ (more C_4 like; Alenazi *et al.*, 2023). In the C_3+C_4 populations, there is a range from 0.05 to 0.29, with variation between subclades (Fig. 5a). We identified five regions of the genome correlated with IBSF, each on a different chromosome (Tables 1, S3; Fig. 5). Expression was detected in mature leaves for 62% of the 65 genes located in the five regions, with no consistent differential expression between photosynthetic types in mature leaves (Dunning *et al.*, 2019a), although two are on average more highly expressed in the C_3 vs C_4 accessions (*ASEM_AUS1_21119* Bonferroni-adjusted *P*-value = 0.073; *ASEM_AUS1_17094* Bonferroni-adjusted *P*-value = <0.001). Five out of the 65 genes were also evolving under strong positive selection with a dN/dS ratio (ω) > 1 using the one-ratio model (Table S4). The annotated genes in the correlated regions of the genome have a variety of functions (Tables 1, S4), including loci directly connected to the response to light stress (*Ferulate 5-Hydroxylase 1*, *FAH1* (*ASEM_AUS1_36251*) (Maruta *et al.*, 2014)) and leaf development (*GATA transcription factor 19*, *GAT19* (*ASEM_AUS1_21136*), *CCR4-NOT transcription complex subunit 11*, *CNOT11*

New Phytologist

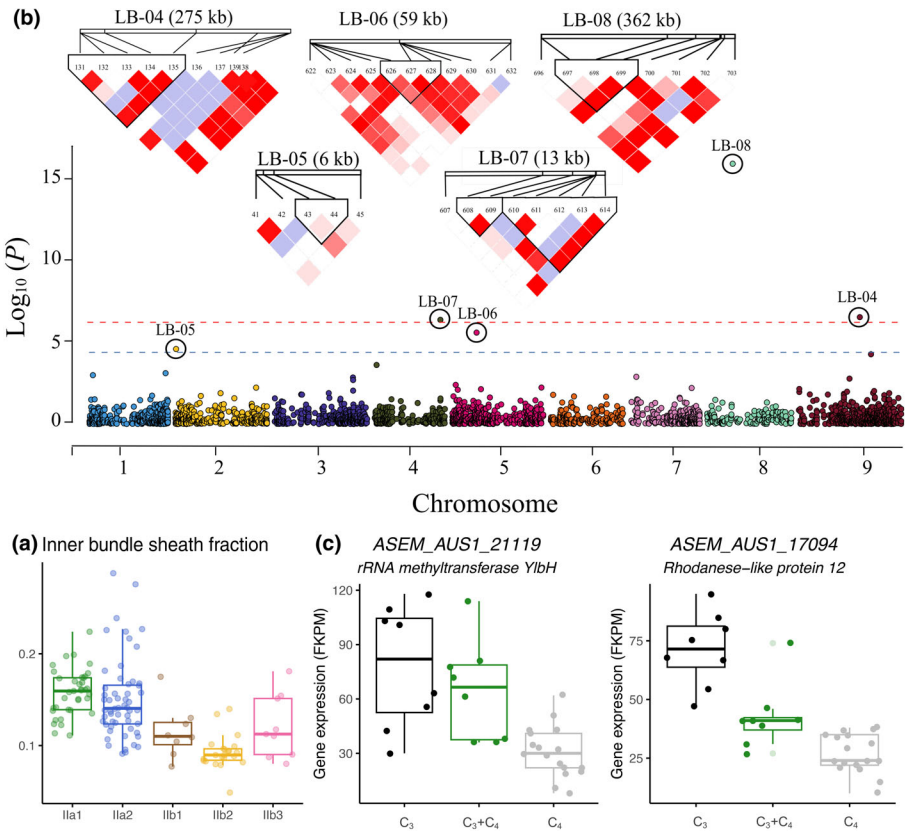

Fig. 5 Genetic variation associated with inner bundle sheath fraction (IBSF) in the C_3+C_4 *Alloteropsis semialata*. (a) The boxplot shows the IBSF variation for each of the C_3+C_4 subclades. The box indicates the median value and the interquartile range, and the whiskers represent 1.5× the interquartile range. (b) A Manhattan plot showing the results of a genome-wide association study (GWAS) for IBSF. The blue and red dotted lines indicate Bonferroni corrected *P*-values of 0.05 and 0.001, respectively. Significant single-nucleotide polymorphisms (SNPs) are labelled with a block ID. Heatmap of pairwise linkage disequilibrium (LD) between markers surrounding each significant SNP, ranging from white indicating low LD (LOD < 2 and D′ < 1) to bright red indicating strong LD (LOD ≥2 and D′ = 1). (c) Boxplot of gene expression for the candidate genes that are significantly differentially expressed. The boxes show the median value and the interquartile range, and the whiskers represent 1.5× the interquartile range.

(*ASEM_AUS1_25789*) & *GRF1-interacting factor 1*, *GIF1* (*ASEM_AUS1_21151*)) (Sarowar *et al.*, 2007; Zhang *et al.*, 2018; An *et al.*, 2020).

Discussion

Alloteropsis semialata has C_3, C_3+C_4, and C_4 genotypes that recently diverged, and it is therefore a useful model to study the initial steps leading to the establishment of the C_4 phenotype since these modifications are not conflated with other changes that accumulate over time (Pereira *et al.*, 2023), and its emergence in this species provided an immediate demographic advantage (Sotelo *et al.*, 2024). Here, we estimate the heritability and identify regions of the genome correlated with variation in both the stable carbon isotope ratio ($\delta^{13}C$) and leaf anatomical traits known to influence $\delta^{13}C$ from field-based measurements (Alenazi *et al.*, 2023). Despite a relatively modest sample size (n = 420 for $\delta^{13}C$; n = 132 for leaf anatomy), we identified regions of the genome significantly associated with these traits, which may indicate that the genetic architecture of C_4 evolution in *A. semialata* is relatively simple, although a broader study may identify additional loci. At present, functional validation of the candidate loci is not possible in *Alloteropsis*, as no proven stable transformation system has been established (Pereira *et al.*, 2023). This is something that would greatly advance the utility of *Alloteropsis semialata* as a model system to study C_4 evolution in the future.

Genetic basis of the carbon isotope ratio ($\delta^{13}C$) in *A. semialata*

Using linked phenotype and genotype information for 420 *A. semialata* individuals, we identified three associated regions of the genome, containing eight protein-coding genes (Fig. 3). The underlying differences in the $\delta^{13}C$ between photosynthetic types are driven by C_4 plants evolving to fix carbon with the PEPC enzyme rather than Rubisco. However, genes encoding PEPC were not detected in the associated regions identified in our GWAS. This absence could be due to variation in the specific PEPC gene copy used for C_4 in the different individuals masking the signal, with up to five different versions known to be used by different *A. semialata* populations (Dunning *et al.*, 2017). Among these five copies, three were laterally acquired (Christin *et al.*, 2012), complicating the matter further as they appear as large structural variants inserted randomly into the genome (Dunning *et al.*, 2019b) and are only present in a subset of individuals (Raimondeau *et al.*, 2023). However, based on the annotations of the genes in the associated regions, we did identify candidate genes with functions potentially associated with the $\delta^{13}C$, the most promising of which include those co-expressed with Rubisco (*SLP1* (*ASEM_AUS1_34302*)), the activation of the NADP-ME C_4 decarboxylating enzyme (*RIPK* (*ASEM_AUS1_34305*)), the development of C_4 'Kranz' anatomy (*SCL9* (*ASEM_AUS1_29467*)), and the suppression of nonphotochemical quenching (*SOQ1* (*ASEM_AUS1_14480*)).

SLP1 encodes a Shewanella-like protein phosphatase 1, an ancient chloroplast phosphatase that is generally more highly expressed in photosynthetic tissue (Kutuzov & Andreeva, 2012; Johnson *et al.*, 2020). In *Arabidopsis thaliana*, it is co-expressed with a number of photosynthetic genes (including all of the Calvin cycle enzymes and Rubisco activase) and it is predicted to play a role in the light-dependent regulation of chloroplast function (Kutuzov & Andreeva, 2012). In *A. semialata*, *SLP1* is significantly more highly expressed in the C_3 individuals compared with the other photosynthetic types. This greater expression in C_3 individuals could indicate a higher Calvin cycle activity at the whole leaf level; meanwhile, in the C_3+C_4 and C_4 individuals, its expression would be increasingly restricted to the IBS tissue. Subdivision of the light signaling networks is one of the key steps in the partitioning of photosynthesis across tissue types in C_4 species (Hendron & Kelly, 2020), and *SLP1* is potentially one of the regulators of this key innovation in *A. semialata*.

RIPK is an enzyme that plays a role in disease resistance and plant immunity (Liu *et al.*, 2011), but has pleiotropic effects. In *A. thaliana*, RIPK directly phosphorylates NADP-ME2 (*AT5G11670*) to enhance its activity and increase cytosolic NADPH concentrations (Wu *et al.*, 2022). In C_4 species, CO_2 is initially fixed in the mesophyll by CA and PEPC before being transported to an internal leaf compartment and released for Rubisco to assimilate through the Calvin cycle. Preliminary studies in *A. semialata* concluded that NADP-ME was the predominant decarboxylating enzyme, although its activity varied with temperature (Frean *et al.*, 1983). Subsequent transcriptome work showed that NADP-ME expression (specifically the *nadpme-1P4* gene that is a many-to-many ortholog of *AT5G11670*) has a mean expression level four times higher in C_4 and C_3+C_4 individuals (mean = 300 RPKM; SD = 235) than in C_3 plants (mean = 75 RPKM; SD = 32), although this difference is not always consistent between populations (Dunning *et al.*, 2019a). The other decarboxylating enzyme commonly used by C_4 *Alloteropsis* is phosphoenolpyruvate carboxykinase (PCK), but like PEPC, a C_4 copy of PCK was also laterally acquired (Christin *et al.*, 2012), complicating its identification in a GWAS analysis because it is absent in the C_3 (Dunning *et al.*, 2019b).

SCL9 belongs to the GRAS gene family of transcription factors that regulate plant development (Hirsch & Oldroyd, 2009). This multigene family includes two known C_4 Kranz anatomy regulators identified in maize, *SHORTROOT* (Slewinski *et al.*, 2014) and *SCARECROW* (Slewinski *et al.*, 2012). Orthologous *SCARECROW* (*SCR*) genes have divergent functions, being recruited for distinct roles in leaf development within maize, rice, and *A. thaliana* (Hughes & Langdale, 2022). In addition to its influence on leaf anatomy, SCR is also required for maintaining photosynthetic capacity in maize (Hughes & Langdale, 2020). The correlation of the *SCARECROW-LIKE SCL9* gene with the strength of the C_4 cycle in *A. semialata* may indicate that convergence in C_4 phenotypes are a result of the parallel recruitment of GRAS transcription factors between species, although there is divergence in the specific loci recruited for this purpose.

SOQ1 is a chloroplast-localized thylakoid membrane protein that regulates nonphotochemical quenching in *A. thaliana* (Brooks *et al.*, 2013; Duan *et al.*, 2023). In full sunlight, plants absorb more light energy than they can process, which can ultimately result in the generation of free radicals that damage the photosynthetic apparatus (Müller *et al.*, 2001). To overcome this, plants have evolved nonphotochemical quenching, which enables them to dissipate the excess energy as heat. This problem is potentially exacerbated in C_4 species, which typically grow in high-light conditions compared with their C_3 counterparts (Sage & Monson, 1999). Preliminary evidence indicates that C_4 species exhibit a significantly faster and greater nonphotochemical quenching relaxation than their C_3 relatives, including between photosynthetic types in *A. semialata* (Acre Cubas, 2023). *SOQ1* may therefore play a direct role in regulating differences in the nonphotochemical quenching responses among *A. semialata* photosynthetic types, and it may represent a good candidate gene to target for reduced photoinhibition associated with fluctuating light conditions in crops (Long *et al.*, 1994).

The genetic basis of C_4 leaf anatomy

In *A. semialata*, the IBS is the site of C_4 photosynthesis, and three leaf anatomical variables linked to the proliferation of this tissue explain the strength of the C_4 cycle ($\delta^{13}C$) in the C_3+C_4 intermediate individuals: IBSW, BSD, and IBSF (Alenazi *et al.*, 2023). IBSW has the lowest heritability (h^2 = 0.06 (SE = 0.06)), and we failed to identify any significant SNPs correlated with this phenotype in our GWAS. This absence of significant genetic factors contributing to the trait may indicate that IBSW has a complex genetic architecture or high phenotypic plasticity. In *A. semialata*, $\delta^{13}C$ is largely genetically based as it is highly heritable after population structure has been accounted for (h^2 = 0.75 (SE = 0.06)), and field-based differences are preserved in a common environment (Lundgren *et al.*, 2016). However, slight variation in $\delta^{13}C$ can still be caused by environmental effects on water use efficiency (Farquhar & Richards, 1984). The previously observed correlation of field-based IBSW measurements with $\delta^{13}C$ may potentially arise from such environmental-induced plasticity (Alenazi *et al.*, 2023). For example, bundle sheath cells in wheat have a larger diameter (more C_4-like) under drought conditions (David *et al.*, 2017).

Plasmodesmata and reduced distance between bundle sheaths

We identified two regions of the genome associated with BSD that contain a single protein-coding gene. This gene is *GSL8* (*ASEM_AUS1_16831*), a member of the GSL family that encodes enzymes synthesizing callose. *GSL8* plays an important role in tissue-level organization (Chen *et al.*, 2009), including stomatal (Guseman *et al.*, 2010) and leaf vein patterning (Linh & Scarpella, 2022). Mutants of *GSL8* in *A. thaliana* formed networks of fewer veins in their leaves (Linh & Scarpella, 2022). This change in venation is mediated by the aperture of plasmodesmata, channels through cell walls that connect neighboring cells (Paterlini, 2020; Band, 2021), which is regulated by GSL8 (Saatian *et al.*, 2018; Linh & Scarpella, 2022). Normal vein

patterning is reliant on an auxin hormone signal travelling through these plasmodesmata, and any interference of this signal disrupts leaf vein development (Linh & Scarpella, 2022). GSL8 might play a role in strengthening the C_4 cycle in *A. semialata* by reducing the distance between bundle sheaths through modulation of the auxin signal. The transition to being fully C_4 in *A. semialata* is also correlated with the presence of minor veins, which reduces both the number of mesophyll cells and the distance between bundle sheaths in C_3+C_4 in comparison with C_3 populations (Lundgren *et al.*, 2019). Therefore, GSL8 may play a pleiotropic role in the strengthening of C_4 photosynthesis in *A. semialata* by increasing both the proportion of bundle sheath tissue in the leaf, and the connectivity between the two distinct cell types required to complete the cycle. The $\delta^{13}C$ GWAS using the C_4 and C_3+C_4 (which differ in the presence of minor veins; Lundgren *et al.*, 2019) interestingly identified a paralog of a gene recently shown to specify vein rank in maize (Vlad *et al.*, 2024), and potentially, *WIP2* has been co-opted for a similar function in *A. semialata*.

The genetic basis of the inner bundle sheath fraction in *A. semialata*

The inner bundle sheath fraction has the highest heritability of all the three leaf anatomy measures used (h^2 = 0.22 (SE = 0.04)). Since it is a composite trait, it is more likely to be influenced by multiple developmental processes. Our GWAS identified five regions of the genome significantly associated with IBSF, containing 65 predicted protein-coding genes. Interestingly, we found a number of genes associated with leaf development that could play a role in the development of C_4 leaf architecture. These include homologs of genes that alter leaf area and vascular development (*GAT19* (*ASEM_AUS1_21136*)) (An *et al.*, 2020), leaf thickness (*CNOT11* (*ASEM_AUS1_25789*)) (Sarowar *et al.*, 2007), and leaf width by regulating meristem determinacy (*GIF1* (*ASEM_AUS1_21151*)) (Zhang *et al.*, 2018). *GIF1* (also called *ANGUSTIFOLIA3*) is perhaps the most interesting of these genes, since it is expressed in the mesophyll cells of leaf primordium and can influence the proliferation of other clonally independent leaf cells (e.g. epidermal cells (Kawade *et al.*, 2013)). The numerous regulators of leaf development identified in the GWAS point to an interacting balance of growth regulators to increase the proportion of bundle sheath tissue within the leaf for C_4 photosynthesis.

There are other genes in these regions with a diverse set of functions, although it is unclear how they could modulate IBSF, including genes associated with light stress and lignin biosynthesis. *FAH1* encodes ferulic acid 5-hydroxylase (F5H) 1, a cytochrome P450 protein that, when disrupted, reduces anthocyanin accumulation under photooxidative stress (Maruta *et al.*, 2014) and is more highly expressed in the C_3 (mean RPKM = 7.38; SD = 5.29) than other photosynthetic types (mean RPKM = 1.00; SD = 2.10; Table S4). These loci could also play a role in C_4 photosynthesis, although most likely they might just be in close physical linkage.

Conclusion

C_4 photosynthesis is a complex trait that requires the rewiring of metabolic gene networks and alterations to the internal leaf anatomy. We investigated the genetic basis of these key innovations in *Alloteropsis semialata*, which has recently diverged C_3, C_3+C_4 intermediate, and C_4 phenotypes. We performed a GWAS that identified regulators of C_4 decarboxylation enzymes (*RIPK*), nonphotochemical quenching (*SOQ1*), and several genes involved in tissue-level organization and leaf development (e.g. *SCL9*, *GSL8*, and *GIF1*). Interestingly, these tend to come from the same gene families as the previously identified C_4 leaf anatomy regulators in other species. This parallel recruitment appears to mirror the pattern observed in the core metabolic enzymes, with the paralog recruited for the C_4 function depending on its ancestral expression pattern and catalytic properties (Wang *et al.*, 2009; Hibberd & Covshoff, 2010; Christin *et al.*, 2013, 2015; Aubry *et al.*, 2014; Emms *et al.*, 2016; Moreno-Villena *et al.*, 2018). Thus, the easiest path to C_4 leaf anatomy would be context-dependent, which likely has implications for engineering C_4 anatomy in C_3 species.

Acknowledgements

The authors extend their appreciation to the Deanship of Scientific Research at Northern Border University, Arar, KSA, for funding this research work (project no.: NBU-SAFIR-2024). LP is supported by a Natural Environment Research Council (grant no.: NE/V000012/1). PAC was funded by a Royal Society University Research Fellowship (grant no.: URF\R\180022). LTD is funded by a NERC fellowship (grant no.: NE/T011025/1).

Competing interests

None declared.

Author contributions

ASA, LP, P-AC, CPO and LTD designed the study. ASA conducted the experimental work and generated the phenotype data. ASA, LP and LTD analyzed the data. All authors interpreted the results and helped write the manuscript. ASA and LP contributed equally to this work.

ORCID

Ahmed S. Alenazi https://orcid.org/0000-0003-4105-2539
Pascal-Antoine Christin https://orcid.org/0000-0001-6292-8734
Luke T. Dunning https://orcid.org/0000-0002-4776-9568
Colin P. Osborne https://orcid.org/0000-0002-7423-3718
Lara Pereira https://orcid.org/0000-0001-5184-8587

Data availability

Most *Alloteropsis semialata* genomic and phenotypic data were previously published, and the source details are specified in

Tables S1 and S2. The additional phenotype data generated here are available in Table S2.

References

Acre Cubas L. 2023. *A comparative analysis of C_3 and C_4 photosynthesis under dynamic light conditions*. PhD thesis, Cambridge University, UK.

Adachi S, Stata M, Martin DG, Cheng S, Liu H, Zhu XG, Sage RF. 2023. The evolution of C_4 photosynthesis in *Flaveria* (Asteraceae): insights from the *Flaveria linearis* complex. *Plant Physiology* **191**: 233–251.

Alenazi AS, Bianconi ME, Middlemiss E, Milenkovic V, Curran EV, Sotelo G, Lundgren MR, Nyirenda F, Pereira L, Christin P-A *et al.* 2023. Leaf anatomy explains the strength of C_4 activity within the grass species *Alloteropsis semialata*. *Plant, Cell & Environment* **8**: 2310–2322.

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research* **9**: 1655–1664.

An Y, Zhou Y, Han X, Shen C, Wang S, Liu C, Yin W, Xia X. 2020. The GATA transcription factor GNC plays an important role in photosynthesis and growth in poplar. *Journal of Experimental Botany* **71**: 1969–1984.

Aubry S, Kelly S, Kümpers BMC, Smith-Unna RD, Hibberd JM. 2014. Deep evolutionary comparison of gene expression identifies parallel recruitment of *trans*-factors in two independent origins of C_4 photosynthesis. *PLoS Genetics* **10**: e1004365.

Band LR. 2021. Auxin fluxes through plasmodesmata. *New Phytologist* **231**: 1686–1692.

Barrett JC, Fry B, Maller J, Daly MJ. 2005. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **2**: 263–265.

Bender MM. 1968. Mass spectrometric studies of carbon 13 variations in corn and other grasses. *Radiocarbon* **10**: 468–472.

Berardini TZ, Reiser L, Li D, Mezheritsky Y, Muller R, Strait E, Huala E. 2015. The Arabidopsis information resource: making and mining the "gold standard" annotated reference plant genome. *Genesis* **53**: 474–485.

Bianconi ME, Dunning LT, Curran EV, Hidalgo O, Powell RF, Mian S, Leitch IJ, Lundgren MR, Manzi S, Vorontsova MS *et al.* 2020. Contrasted histories of organelle and nuclear genomes underlying physiological diversification in a grass species. *Proceedings of the Royal Society B: Biological Sciences* **287**: 20201960.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **15**: 2114–2120.

Brooks MD, Sylak-Glassman EJ, Fleming GR, Niyogi KK. 2013. A thioredoxin-like/β-propeller protein maintains the efficiency of light harvesting in *Arabidopsis*. *Proceedings of the National Academy of Sciences, USA* **110**: E2733–E2740.

Cerros-Tlatilpa R, Columbus JT. 2009. C_3 photosynthesis in *Aristida longifolia*: implication for photosynthetic diversification in Aristidoideae (Poaceae). *American Journal of Botany* **96**: 1379–1387.

Chen L, Ge B, Casale FP, Vasquez L, Kwan T, Garrido-Martín D, Ecker S. 2016. Genetic drivers of epigenetic and transcriptional variation in human immune cells. *Cell* **5**: 1398–1414.

Chen XY, Liu L, Lee E, Han X, Rim Y, Chu H, Kim SW, Sack F, Kim JY. 2009. The Arabidopsis callose synthase gene *GSL8* is required for cytokinesis and cell patterning. *Plant Physiology* **150**: 105–113.

Christin P-A, Arakaki M, Osborne CP, Edwards EJ. 2015. Genetic enablers underlying the clustered evolutionary origins of C_4 photosynthesis in angiosperms. *Molecular Biology and Evolution* **32**: 846–858.

Christin P-A, Boxall SF, Gregory R, Edwards EJ, Hartwell J, Osborne CP. 2013. Parallel recruitment of multiple genes into C_4 photosynthesis. *Genome Biology and Evolution* **5**: 2174–2187.

Christin PA, Edwards EJ, Besnard G, Boxall SF, Gregory R, Kellogg EA, Hartwell J, Osborne CP. 2012. Adaptive evolution of C_4 photosynthesis through recurrent lateral gene transfer. *Current Biology* **22**: 445–449.

Cui H, Kong D, Liu X, Hao Y. 2014. SCARECROW, SCR-LIKE 23 and SHORT-ROOT control bundle sheath cell fate and function in *Arabidopsis thaliana*. *The Plant Journal* **78**: 319–327.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Durbin R. 2011. The variant call format and VCFtools. *Bioinformatics* **15**: 2156–2158.

David OA, Osonubi O, Olaiya CO, Agbolade JO, Ajiboye AA, Komolafe RJ, Chukwuma DM, Akomolafe GF. 2017. Anatomical response of wheat cultivars to drought stress. *IFE Journal of Science* **19**: 323–331.

Duan S, Dong B, Chen Z, Hong L, Zhang P, Yang Z, Wang H-B, Jin H-L. 2023. HHL1 and SOQ1 synergistically regulate nonphotochemical quenching in Arabidopsis. *Journal of Biological Chemistry* **299**: 104670.

Dunning LT, Lundgren MR, Moreno-Villena JJ, Namaganda M, Edwards EJ, Nosil P, Osborne CP, Christin PA. 2017. Introgression and repeated co-option facilitated the recurrent emergence of C_4 photosynthesis among close relatives. *Evolution* **71**: 1541–1555.

Dunning LT, Moreno-Villena JJ, Lundgren MR, Dionora J, Salazar P, Adams C, Nyirenda F, Olofsson JK, Mapaura A, Grundy IM *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, Papadopulos AS, Hibdige SG, Hidalgo O, Leitch IJ, Baleeiro PC, Ntshangase S, Barker N, Jobson RW. 2022. Hybridisation and chloroplast capture between distinct *Themeda triandra* lineages in Australia. *Molecular Ecology* **31**: 5846–5860.

Dunning LT, Olofsson JK, Parisod C, Choudhury R, Moreno-Villena J, Yang Y, Dionora J, Quick WP, Park M, Bennetzen JL *et al.* 2019b. Lateral transfers of large DNA fragments spread functional genes among grasses. *Proceedings of the National Academy of Sciences, USA* **10**: 4416–4425.

Edwards GE, Ku MS. 1987. Biochemistry of C_3–C_4 intermediates. In: Hatch MD, Boardman, eds. *The biochemistry of plants: a comprehensive treatise, vol. 10*. New York, NY, USA: Academic Press, 275–325.

Emms DM, Covshoff S, Hibberd JM, Kelly S. 2016. Independent and parallel evolution of new genes by gene duplication in two origins of C_4 photosynthesis provides new insight into the mechanism of phloem loading in C_4 species. *Molecular Biology and Evolution* **33**: 1796–1806.

Emms DM, Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology* **1**: 157.

Farquhar GD, Ehleringer JR, Hubick KT. 1989. Carbon isotope discrimination and photosynthesis. *Annual Review of Plant Physiology and Plant Molecular Biology* **40**: 503–537.

Farquhar GD, O'Leary M, Berry J. 1982. On the relationship between carbon isotope discrimination and the intercellular carbon dioxide concentration in leaves. *Australian Journal of Plant Physiology* **2**: 121–137.

Farquhar GD, Richards RA. 1984. Isotopic composition of plant carbon correlates with water-use efficiency of wheat genotypes. *Functional Plant Biology* **11**: 539–552.

Ferguson JN, Fernandes SB, Monier B, Miller ND, Allen D, Dmitrieva A, Schmuker P, Lozano R, Valluru R, Buckler ES *et al.* 2021. Machine learning-enabled phenotyping for GWAS and TWAS of WUE traits in 869 field-grown sorghum accessions. *Plant Physiology* **187**: 1481–1500.

Frean ML, Ariovich D, Cresswell CF. 1983. C_3 and C_4 photosynthetic and anatomical forms of *Alloteropsis semialata* (R. Br.) Hitchcock: 2. A comparative investigation of leaf ultrastructure and distribution of chlorenchyma in the two forms. *Annals of Botany* **51**: 811–821.

Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Rokhsar DS. 2012. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Research* **40**: D1178–D1186.

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? *Plant Cell* **23**: 2087–2105.

Guseman JM, Lee JS, Bogenschutz NL, Peterson KM, Virata RE, Xie B, Kanaoka MM, Hong Z, Torii KU. 2010. Dysregulation of cell-to-cell connectivity and stomatal patterning by loss-of-function mutation in *Arabidopsis chorus* (*glucan synthase-like 8*). *Development* **137**: 1731–1741.

Hatch MD. 1971. The C_4 pathway of photosynthesis. Evidence for an intermediate pool of carbon dioxide and the identity of the donor C_4 dicarboxylic acid. *Biochemical Journal* **125**: 425–432.

Hatch MD. 1987. C_4 photosynthesis: a unique blend of modified biochemistry, anatomy and ultrastructure. *Biochimica et Biophysica Acta (BBA) – Reviews on Bioenergetics* **895**: 81–106.

Hendron RW, Kelly S. 2020. Subdivision of light signaling networks contributes to partitioning of C_4 photosynthesis. *Plant Physiology* **182**: 1297–1309.

Heyduk K, Moreno-Villena JJ, Gilman I, Christin PA, Edwards EJ. 2019. The genetics of convergent evolution: insights from plant photosynthesis. *Nature Reviews Genetics* **20**: 485–493.

Hibberd JM, Covshoff S. 2010. The regulation of gene expression required for C_4 photosynthesis. *Annual Review of Plant Biology* **61**: 181–207.

Hirsch S, Oldroyd GE. 2009. GRAS-domain transcription factors that regulate plant development. *Plant Signaling & Behavior* **4**: 698–700.

Hughes TE, Langdale JA. 2020. *SCARECROW* gene function is required for photosynthetic development in maize. *Plant Direct* **4**: e00264.

Hughes TE, Langdale JA. 2022. SCARECROW is deployed in distinct contexts during rice and maize leaf development. *Development* **149**: dev200410.

Hugin Development Team. 2015. *Hugin – Panorama photo stitcher*. [WWW document] URL https://hugin.sourceforge.io [accessed 1 January 2021].

Johnson JJ, White-Gloria C, Toth R, Labandera AM, Uhrig RG, Moorhead GB. 2020. SLP1 and SLP2: ancient chloroplast and mitochondrial protein phosphatases. In: *Protein phosphatases and stress management in plants: functional genomic perspective*. Cham, Switzerland: Springer, 1–9.

Kawade K, Horiguchi G, Usami T, Hirai MY, Tsukaya H. 2013. ANGUSTIFOLIA3 signaling coordinates proliferation between clonally distinct cells in leaves. *Current Biology* **23**: 788–792.

Kim SA, Cho CS, Kim SR, Bull SB, Yoo YJ. 2018. A new haplotype block detection method for dense genome sequencing data based on interval graph modeling of clusters of highly correlated SNPs. *Bioinformatics* **34**: 388–397.

Kutuzov MA, Andreeva AV. 2012. Prediction of biological functions of *Shewanella*-like protein phosphatases (Shelphs) across different domains of life. *Functional & Integrative Genomics* **12**: 11–23.

Langdale JA, Metzler MC, Nelson T. 1987. The *argentia* mutation delays normal development of photosynthetic cell-types in *Zea mays*. *Developmental Biology* **122**: 243–255.

Langdale JA, Rothermel BA, Nelson T. 1988. Cellular pattern of photosynthetic gene expression in developing maize leaves. *Genes & Development* **1**: 106–115.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**: 357–359.

Linh NM, Scarpella E. 2022. Leaf vein patterning is regulated by the aperture of plasmodesmata intercellular channels. *PLoS Biology* **20**: e3001781.

Liu J, Elmore JM, Lin ZJD, Coaker G. 2011. A receptor-like cytoplasmic kinase phosphorylates the host target RIN4, leading to the activation of a plant innate immune receptor. *Cell Host & Microbe* **9**: 137–146.

Long SP, Humphries S, Falkowski PG. 1994. Photoinhibition of photosynthesis in nature. *Annual Review of Plant Biology* **45**: 633–662.

Lundgren MR, Besnard G, Ripley BS, Lehmann CER, Chatelet DS, Kynast RG, Namaganda M, Vorontsova MS, Hall RC, Elia J *et al.* 2015. Photosynthetic innovation broadens the niche within a single species. *Ecology Letters* **18**: 1021–1029.

Lundgren MR, Christin P-A, 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, Dunning LT, Olofsson JK, Moreno-Villena JJ, Bouvier JW, Sage TL, Khoshravesh R, Sultmanis S, Stata M, Ripley BS *et al.* 2019. C_4 anatomy can evolve via a single developmental change. *Ecology Letters* **22**: 302–312.

Maruta T, Noshi M, Nakamura M, Matsuda S, Tamoi M, Ishikawa T, Shigeoka S. 2014. Ferulic acid 5-hydroxylase 1 is essential for expression of anthocyanin biosynthesis-associated genes and anthocyanin accumulation under photooxidative stress in *Arabidopsis*. *Plant Science* **219–220**: 61–68.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M *et al.* 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20**: 1297–1303.

Monaco MK, Sen TZ, Dharmawardhana PD, Ren L, Schaeffer M, Naithani S, Amarasinghe V, Thomason J, Harper L, Gardiner J *et al.* 2013. Maize metabolic network construction and transcriptome analysis. *The Plant Genome* **1**: 1–12.

Moreno-Villena JJ, Dunning LT, Osborne CP, Christin P-A. 2018. Highly expressed genes are preferentially co-opted for C_4 photosynthesis. *Molecular Biology and Evolution* **35**: 94–106.

Müller P, Li XP, Niyogi KK. 2001. Non-photochemical quenching. A response to excess light energy. *Plant Physiology* **125**: 1558–1566.

O'Leary MH. 1981. Carbon isotope fractionation in plants. *Phytochemistry* **20**: 553–567.

Olofsson JK, Bianconi M, Besnard G, Dunning LT, Lundgren MR, Holota H, Vorontsova MS, Hidalgo O, Leitch IJ, Nosil P *et al.* 2016. Genome biogeography reveals the intraspecific spread of adaptive mutations for a complex trait. *Molecular Ecology* **25**: 6107–6123.

Olofsson JK, Curran EV, Nyirenda F, Bianconi ME, Dunning LT, Milenkovic V, Sotelo G, Hidalgo O, Powell RF, Lundgren MR *et al.* 2021. Low dispersal and ploidy differences in a grass maintain photosynthetic diversity despite gene flow and habitat overlap. *Molecular Ecology* **9**: 2116–2130.

Ortiz EM. 2019. *vcf2phylip v.2.0: convert a VCF matrix into several matrix formats for phylogenetic analysis.*

Paterlini A. 2020. Uncharted routes: exploring the relevance of auxin movement via plasmodesmata. *Biology Open* **9**: 11.

Pereira L, Bianconi ME, Osborne CP, Christin P-A, Dunning LT. 2023. *Alloteropsis semialata* as a study system for C_4 evolution in grasses. *Annals of Botany* **132**: 365–382.

Petricka JJ, Clay NK, Nelson TM. 2008. Vein patterning screens and the *defectively organized tributaries* mutants in *Arabidopsis thaliana*. *The Plant Journal* **56**: 251–263.

Pignon CP, Leakey AD, Long SP, Kromdijk J. 2021. Drivers of natural variation in water-use efficiency under fluctuating light are promising targets for improvement in Sorghum. *Frontiers in Plant Science* **12**: 627432.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Sham PC. 2007. Plink: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics* **3**: 559–575.

Raimondeau P, Bianconi ME, Pereira L, Parisod C, Christin PA, Dunning LT. 2023. Lateral gene transfer generates accessory genes that accumulate at different rates within a grass lineage. *New Phytologist* **240**: 2072–2084.

Saatian B, Austin RS, Tian G, Chen C, Nguyen V, Kohalmi SE, Geelen D, Cui Y. 2018. Analysis of a novel mutant allele of *GSL8* reveals its key roles in cytokinesis and symplastic trafficking in Arabidopsis. *BMC Plant Biology* **18**: 1–17.

Sage RF, Monson RK. 1999. *C_4 plant biology*. San Diego, CA, USA: Academic Press.

Sakai H, Lee SS, Tanaka T, Numa H, Kim J, Kawahara Y, Wakimoto H, Yang C-C, Iwamoto M, Abe T *et al.* 2013. Rice annotation project database (RAP-DB): an integrative and interactive database for rice genomics. *Plant and Cell Physiology* **54**: e6.

Sarowar S, Oh HW, Cho HS, Baek KH, Seong ES, Joung YH, Choi GJ, Lee S, Choi D. 2007. *Capsicum annuum* CCR4-associated factor *CaCAF1* is necessary for plant development and defence response. *The Plant Journal* **51**: 792–802.

Schneider CA, Rasband WS, Eliceiri KW. 2012. NIH Image to ImageJ: 25 years of image analysis. *Nature Methods* **9**: 671–675.

Simpson CJC, Reeves G, Tripathi A, Singh P, Hibberd JM. 2021. Using breeding and quantitative genetics to understand the C_4 pathway. *Journal of Experimental Botany* **73**: 3072–3084.

Slewinski TL, Anderson AA, Price S, Withee JR, Gallagher K, Turgeon R. 2014. Short-root1 plays a role in the development of vascular tissue and Kranz anatomy in maize leaves. *Molecular Plant* **7**: 1388–1392.

Slewinski TL, Anderson AA, Zhang C, Turgeon R. 2012. *Scarecrow* plays a role in establishing Kranz anatomy in maize leaves. *Plant and Cell Physiology* **53**: 2030–2037.

Smith BN, Brown WV. 1973. The Kranz syndrome in the Gramineae as indicated by carbon isotopic ratios. *American Journal of Botany* **60**: 505–513.

Smith BN, Epstein S. 1971. Two categories of $^{13}C/^{12}C$ ratios for higher plants. *Plant Physiology* **47**: 380–384.

Sotelo G, Gamboa S, Dunning LT, Christin P-A, Varela S. 2024. C_4 photosynthesis provided an immediate demographic advantage to populations of the grass *Alloteropsis semialata*. *New Phytologist* **242**: 774–785.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**: 1312–1313.

Stata M, Sage TL, Sage RF. 2019. Mind the gap: the evolutionary engagement of the C_4 metabolic cycle in support of net carbon assimilation. *Current Opinion in Plant Biology* **49**: 27–34.

Strigens A, Schipprack W, Reif JC, Melchinger AE. 2013. Unlocking the genetic diversity of maize landraces with doubled haploids opens new avenues for breeding. *PLoS ONE* **2**: e57234.

Vlad D, Zaidem M, Perico C, Sedelnikova O, Bhattacharya S, Langdale JA. 2024. The WIP6 transcription factor *TOO MANY LATERALS* specifies vein type in C_4 and C_3 grass leaves. *Current Biology* **34**: 1670–1686.

Von Caemmerer S. 1992. Carbon isotope discrimination in C_3–C_4 intermediates. *Plant, Cell & Environment* **15**: 1063–1072.

Voznesenskaya EV, Koteyeva NK, Chuong SDX, Ivanova AN, Barroca J, Craven LA, Edwards GE. 2007. Physiological, anatomical and biochemical characterisation of photosynthetic types in genus *Cleome* (Cleomaceae). *Functional Plant Biology* **34**: 247–267.

Wang X, Gowik U, Tang H, Bowers JE, Westhoff P, Paterson AH. 2009. Comparative genomic analysis of C_4 photosynthetic pathway evolution in grasses. *Genome Biology* **10**: R68.

Wu B, Li P, Hong X, Xu C, Wang R, Liang Y. 2022. The receptor-like cytosolic kinase RIPK activates NADP-malic enzyme 2 to generate NADPH for fueling ROS production. *Molecular Plant* **5**: 887–903.

Yang J, Lee SH, Goddard ME, Visscher PM. 2011. GCTA: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics* **88**: 76–82.

Yin L, Zhang H, Tang Z, Xu J, Yin D, Zhang Z, Liu X. 2021. rMVP: a memory-efficient, visualization-enhanced, and parallel-accelerated tool for genome-wide association study. *Genomics, Proteomics & Bioinformatics* **4**: 619–628.

Zhang D, Sun W, Singh R, Zheng Y, Cao Z, Li M, Lunde C, Hake S, Zhang Z. 2018. *GRF-interacting factor1* regulates shoot architecture and meristem determinacy in maize. *Plant Cell* **30**: 360–374.

Supporting Information

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

Dataset S1 Maximum likelihood phylogeny of all accessions used in this study.

Dataset S2 Orthogroup phylogenies.

Fig. S1 Distribution of linkage block sizes in the *Alloteropsis semialata* genome.

Fig. S2 Density and *Q*–*Q* plots for the studied traits.

Fig. S3 Manhattan plot showing the results of a Genome-Wide Association Study for $\delta^{13}C$ using various subdivisions of samples based on photosynthetic type.

Fig. S4 Manhattan plot showing the results of a Genome-Wide Association Study for inner bundle sheath width.

Table S1 Details of samples used in the genome-wide association study.

Table S2 Leaf anatomy measurements for the C_3+C_4 *Alloteropsis semialata*.

Table S3 Summary of significant regions detected.

Table S4 Summary of genes in significant regions.

Please note: Wiley is not responsible for the content or functionality of any Supporting Information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.