Genetic diversity of C4 photosynthesis pathway genes in Sorghum bicolor (L.)

Yongfu Tao (y.tao1@uq.edu.au)
University of Queensland https://orcid.org/0000-0001-9096-7407
Barbara George-Jaeggli
University of Queensland
Marie Bouteille-Pallas
University of Queensland
Shuaishuai Tai
BGI
Alan Cruickshank
Queensland Department of Agriculture and Fisheries
David Jordan
University of Queensland
Emma Mace
University of Queensland

Research article

Keywords: Sorghum, genetic diversity, SNPs, C4 pathway, domestication

Posted Date: October 12th, 2019

DOI: https://doi.org/10.21203/rs.2.15980/v1

Version of Record: A version of this preprint was published at Genes on July 16th, 2020. See the published version at https://doi.org/10.3390/genes11070806.

Abstract

Background C4 photosynthesis has evolved in over 60 different plant taxa and is an excellent example of convergent evolution. Plants using the C4 photosynthetic pathway have an efficiency advantage, particularly in hot and dry environments. They account for 23% of global primary production and include some of our most productive cereals. While previous genetic studies comparing phylogenetically related C3 and C4 species have elucidated the genetic diversity underpinning the C4 photosynthetic pathway, no previous studies have described the genetic diversity of the genes involved in this pathway within a C4 crop species. Enhanced understanding of the allelic diversity and selection signatures of genes in this pathway may present opportunities to improve photosynthetic efficiency, and ultimately yield, by exploiting natural variation.

Results Here, we present the first genetic diversity survey of 8 known C4 gene families in an important C4 crop, Sorghum bicolor (L.) Moench using sequence data of 48 genotypes covering wild and domesticated sorghum accessions. Average nucleotide diversity of C4 gene families varied more than 20-fold from the NADP-MDH gene family ($\theta\pi$ =0.2×10-3) to the PPDK gene family ($\theta\pi$ = 5.21×10-3). Genetic diversity of C4 genes was reduced by 22.43% in cultivated sorghum compared to wild and weedy sorghum, indicating that the group of wild and weedy sorghum may constitute an untapped reservoir for alleles related to the C4 photosynthetic pathway. A SNP-level analysis identified purifying selection signals on C4 PPDK and CA genes, and balancing selection signals on C4 PPDK-RP and PEPC genes. Allelic distribution of these C4 genes was consistent with selection signals detected.

Conclusions Domestication of sorghum has reshaped diversity of C4 pathway. A better understanding of the genetic diversity of this pathway in sorghum paves the way for mining the natural allelic variation for the improvement of photosynthesis.

Background

C_4 photosynthesis has independently evolved in more than 60 different plant taxa [1]. The main driver for this convergent evolution is the tendency of Ribulose–1,5-bisphosphate carboxylase (Rubisco), which catalyses the net fixation of carbon dioxide (CO_2), to also act as an oxygenase in the presence of oxygen (O_2). This reaction produces toxic phosphoglycolate which has to be converted to useful metabolites requiring substantial metabolic energy [2]. This wasteful use of CO_2 is termed photorespiration. Photorespiration becomes a major constraint of photosynthesis in situations where CO_2 to O_2 ratios are low and temperatures are high. The evolution of C_4 photosynthesis coincided with declining atmospheric CO_2 concentrations [1, 3] as a mechanism to avoid photorespiration by concentrating CO_2 in the vicinity of Rubisco.

In the majority of C_4 plants, this is achieved via spatial separation of the initial CO_2 fixation and the Calvin-Benson-Bassham cycle in two different cell types, most often mesophyll cells and bundle sheath cells [4]. CO_2 concentration in C_4 bundle sheath cells are up to 10-fold higher than those found in C_3 mesophyll cells [5]. At higher temperatures, C_4 photosynthesis is not only more efficient compared with C_3 photosynthesis in terms of reducing energy losses from photorespiration, but due to the improved efficiency of this pathway, it renders plants more nitrogen and water-use efficient [6, 7]. C_4 plants are more productive than C_3 plants in areas of high light intensities, warm temperatures and low rainfall, such as the sub-tropical and tropical areas around the globe.

Many of the major crops that originated from and are still grown in the warm and dry regions of the world, such as maize, sorghum, millet, sugarcane, *miscanthus* and switchgrass, use the C_4 pathway [8]. C_4 crops account for an estimated 23% of global primary production [9]. Improved photosynthetic capacity has been suggested as the next frontier in lifting crop productivity [10]. The C_4 photosynthesis pathway is a good starting point to improve photosynthetic capacity and resource efficiency in crop plants. Attempts are currently being undertaken to integrate characteristics of the C_4 pathway into C_3 crops [6, 11–13].

However, possibly due to multiple independent evolution of C_4 photosynthesis in different plant taxa [1], large variation also exists among C_4 species in terms of the biochemical pathway. It has long been known that three major biochemical subtypes - NADP-ME, NAD-ME and PCK/NAD-ME - exist among C_4 species [14]. More recently, it has been suggested that mixtures among them exist [15] and that the subtypes vary in their performance under different environmental conditions, e.g. low light [16]. Especially among the grasses, which all of the C_4 cereals belong to, differences in pathway and performance are likely to exist, as C_4 photosynthesis has evolved at least 25 times in this group of plants [17]. Exploring such variation may provide avenues to further improve C_4 photosynthetic efficiency [8].

Sorghum is a NADP-ME subtype C_4 crop well-known for its adaption to drought and high temperatures. It provides staple food for over 500 million people in the semi-arid tropics of Africa and Asia; in addition to being an important source of feed, fibre and biofuel. Due to these characteristics, it is expected to play an increasingly important role in meeting the challenges of feeding the world's growing population under the threat of global warming. Substantial variation in photosynthesis and related traits has been revealed in sorghum [18–22], however, the genetic basis of this variation has not yet been studied.

The recent assembly of whole-genome sequences for a wide range of wild and cultivated sorghum species [23–25], provides an excellent opportunity to explore genetic diversity of genes related to the C_4 photosynthetic pathway. Several high-throughput comparative transcriptomics and evolutionary studies using C_3 and C_4 phylogenetically related species and cell-specific gene expression have elucidated the key genes and regulatory networks that underpin the C_4 photosynthetic pathway [4, 26–34]. In the present study, we explored the genetic variation in genes that have previously been identified as core C_4 genes, mined their allelic diversity and investigated signatures of selection during domestication in sorghum.

Results

Nucleotide diversity of core C_4 gene families in sorghum

Based on 9 genes corresponding to 8 core C_4 enzymes in sorghum, 18 homologous genes were identified across the sorghum genome. In total, 5 CA genes, 2 NADP-MDH genes, 5 NAPD-ME genes, 6 PEPC genes, 3 PPCK genes, 2 PPDK genes, 3 PPDK-RP genes and 1 rbcS gene were identified (Table 1). Nucleotide diversity ($\theta\pi$) of these 27 genes was investigated using sequence data of 48 genotypes covering wild and weedy, and cultivated sorghum [25]. A total number of 4,183 single nucleotide polymorphisms (SNPs) were identified in these 27 genes with 521 SNPs located in coding sequence (CDS) regions (Table 1). These C_4 gene families displayed an average overall nucleotide diversity of $\theta\pi = 2.09\times10^{-3}$, which is comparable to that of 130 housekeeping genes ($\theta\pi = 1.97\times10^{-3}$) [25]. Nucleotide diversity varied dramatically among the C_4 gene families, with the NADIP-MDH genes displaying the lowest levels of diversity across all genotypes (average $\theta\pi = 0.25\times10^{-3}$), followed by NADP-ME genes ($\theta\pi = 0.93\times10^{-3}$), PPCK genes ($\theta\pi = 1.20\times10^{-3}$), PEPC genes ($\theta\pi = 2.11\times10^{-3}$), CA ($\theta\pi = 2.26\times10^{-3}$) and PPDK-RP ($\theta\pi = 2.96\times10^{-3}$), while PPDK genes showed the highest level of diversity ($\theta\pi = 5.21\times10^{-3}$) (Table 2, Figure 2A). The only gene encoding rbcS, *Sobic.005G042000,* had high genetic diversity with $\theta\pi = 4.32\times10^{-3}$across all 48 genotypes, 5.72×10^{-3}in the wild and weedy group and 3.03×10^{-3}in the cultivated group.

Mixed trends were found when comparing C_4 genes with non-C_4 isoforms in each gene family with the average overall genetic diversity of C_4 genes being comparable to that of their non-C_4 counterpart (Table 2). The C_4 PPDK-RP gene (*Sobic.007G166300)* and C_4 NADP-MDH gene (*Sobic.002G324400)* had an overall $\theta\pi$ which was 161.76% and 79.85% higher than their non-C_4 isoforms, respectively, whereas the $\theta\pi$ of the C_4 PPDK gene (*Sobic.009G132900)* was 75.16% lower than that of the non-C_4 PPDK isoform. Nucleotide diversity of C_4 genes in the other gene families was within the range of variation of their non-C_4 isoforms.

Genetic diversity across C_4 gene families was reduced during sorghum domestication. Averaged across all C_4 gene families genetic diversity was reduced by 22.44% in the domesticated compared with the wild and weedy group and when just the 9 core C_4 genes were considered, the reduction was 22.98%. However, the reduction of genetic diversity during domestication in C_4 genes was not significantly different from that in housekeeping genes (Table S2). Among the 27 genes, *Sobic.003G292400,* a non-C_4 NADP-ME isoform, exhibited the most severe reduction in genetic diversity, with a reduction of 98.23%. The C_4 version of that gene, the NADP-ME gene (*Sobic.003G036200)*,, showed the greatest loss of genetic diversity (51.89%) among the C_4 genes, with a Fst between the cultivated and wild and weedy groups of 0.06 (Figure 2B). In contrast, another non-C_4 isoform of NADP-ME (*Sobic.009G069600)*,, a non-C_4 isoform of PPCK (*Sobic.006G148300)* and a non-C_4 CA isoform (*Sobic.003G234600)*,, showed a more than 2- fold increase in genetic diversity in the cultivated group.

Identification of selection signals during domestication across the 27 genes

The selection signature of these C_4 gene families was firstly investigated at the gene level. Based on thresholds of genome-wide rankings described in Mace et al. [25], only one gene (*Sobic.001G326900,* non-C_4 PPDK isoform) was identified as being under balancing selection during sorghum domestication, while no gene was identified as being under purifying selection (Table 1). Subsequent to this, a higher resolution detection of selection signature was conducted at the SNP level using the CDS of the 27 genes. Among 521 SNPs across 27 CDS, 176 were non-synonymous. The number of non-synonymous SNPs within genes varied from 19 in the non-C_4 PPDK-RP isoform (*Sobic.002G324700)* to 0 in the C_4 PPDK (*Sobic.009G132900)*.. The C_4 PEPC gene (*Sobic.010G160700)* had the highest number of non-synonymous SNPs (9) among the 9 C_4 genes (Table 1).

Based on the SNP-level analysis, 24 SNPs across 8 genes were identified as being under purifying selection including 7 non-synonymous SNPs in 6 genes (Table S3). Genes with SNPs under purifying selection included two C_4 isoforms, PPDK (*Sobic.009G132900)* and CA (*Sobic.003G234200)*,, three of 4 non-C_4 NADP-ME (*Sobic.003G280900, Sobic.003G292400, Sobic.009G069600)*,, both two non-C_4 PPDK-RP (*Sobic.002G324500, Sobic.002G324700)* and a non-C_4 PEPC gene (*Sobic.007G106500)*.. Among the 2 C_4 genes with SNPs under selection, *Sobic.009G132900* had 3 synonymous SNPs under purifying selection, while *Sobic.003G234200* had a non-synonymous SNP under purifying selection.

A total of 60 SNPs across 8 genes were identified as being under balancing selection, 7 of which were non-synonymous SNPs distributed across 2 genes (Table S4). The non-C_4 PPDK (*Sobic.001G326900)* had 24 SNPs under balancing selection including 5 non-synonymous SNPs, and additionally had an overall gene-level signature of balancing selection based on the previous analysis. Two C_4 isoforms, PPDK-RP (*Sobic.002G324400)* and PEPC (*Sobic.010G160700)*,, were identified with 3 and 2 SNPs under balancing selection, respectively, although none of them were non-synonymous SNPs. Two non-C_4 PEPC (*Sobic.003G100600, Sobic.004G106900)* were identified with SNPs under balancing selection, with *Sobic.003G100600* having 21 SNPs including 2 non-synonymous SNPs exhibiting signatures of balancing selection. The other 2 genes with SNPs under balancing selection were a non-C_4 CA isoform, *Sobic.002G230100* and a non-C_4 PPCK isoform, *Sobic.004G219900.*

Allelic variation of core C_4 genes under selection in sorghum

A phylogenetic tree was constructed using CDS of the 27 genes to investigate the genetic relationships between the 48 accessions (Figure S1). The inter- and intra-species distribution of private haplotypes of each gene is detailed in Table S5, with the majority (~90%) of the genes with private inter-species haplotypes from *S. propinquum,* e.g. 4 unique haplotypes were observed for the C_4 isoform of PEPC, with the 2 *S. propinquum* accessions sharing a single private haplotype. To further investigate allelic variation of 4 core C_4 genes with SNPs under selection in sorghum, haplotype networks were constructed using CDS SNPs. Based on 16 SNPs within the CDS of the PPDK gene (*Sobic.009G132900)*,, 8 haplotypes were identified. Five haplotypes were identified in wild and weedy, with 3 being private haplotypes and two of them being maintained in cultivated sorghum; two new haplotypes arose in cultivated sorghum after domestication (Figure 3A). Ten haplotypes of one CA gene (*Sobic.003G234200)* were revealed using 33 SNPs, with 4 distinct haplotypes being characterised by the wild and weedy genotypes, two of which were private haplotypes to this group. The remaining two haplotypes were maintained in cultivated sorghum during domestication, with three new haplotypes arising after domestication (Figure 3B). The loss of wild and weedy haplotypes in cultivated sorghum in these two genes was consistent with the finding that they were under purifying selection.

The PPDK-RP gene (*Sobic.002G324400)* had 22 SNPs in the CDS, based on which 5 haplotypes were identified. Two haplotypes were characterised by the wild and weedy genotypes, with the main wild haplotype maintained and further diversifying into two new haplotypes in the cultivated group (Figure 3C). Based on 28 SNPs in the CDS of the C_4 PEPC gene (*Sobic.010G160700)*,, 4 haplotypes were identified. Wild and weedy genotypes encompassed 3 haplotypes and all of them were maintained in cultivated sorghum (Figure 3D). *S. propinquum* had unique haplotypes across all 4 genes, while the *Sorghum bicolor* race *guinea margaritiferum* shared haplotypes with the wild and weedy genotypes in most cases, indicating a closer relationship with the wild and weedy group.

Discussion

The evolution of C_4 photosynthesis has been studied extensively at the cross-species level with signals of adaptive evolution identified on key genes in the C_4 pathway [27, 33, 35–37]. As the evolution of C_4 photosynthesis is driven by environments characterised by low CO_2 availability, such as hot and dry environments in which CO_2 uptake is limited by stomatal closure, it is likely that within-species adaptive variation also exists. However, to our knowledge, studies of within-species allele diversity and signatures of selection on key genes in the C_4 pathway, have not previously been undertaken.

Knowledge of existing natural variation and levels of genetic diversity is a pre-requisite for the optimisation of C_4 photosynthesis. In this study, we performed the first investigation of the genetic diversity of C_4 gene families within a C_4 species. Substantial variation of nucleotide diversity was observed among these 8 C_4 gene families in sorghum, with the NADP-MDH gene family showing the least diversity and the PPDK gene family showing the greatest diversity. Nine core C_4 genes also exhibited varying degrees of genetic diversity, ranging from θπ values of $5.04\times10^{-3}$and 4.32×10^{-3}in PPDK-RP and rbcS to θπ values of $0.33\times10^{-3}$and 0.67×10^{-3}in NADP-MDH and NADP-ME. However despite such low levels of diversity, non-synonymous SNPs were identified in both NADP-MDH and NADP-ME (Table 1). C_4 PPDK was the only gene which did not contain a non-synonymous SNP, despite its fairly large size (Gene size, 12748bp; CDS, 2847bp), indicating the function of this gene is highly conserved.

Cultivated sorghum was domesticated more than 5 thousand years ago in Africa [38–40]. This artificial selection process has morphologically and physiologically reshaped sorghum to better suit human needs, and also resulted in substantial reduction of genetic diversity genome wide in cultivated sorghum compared with wild and weedy types [41, 42]. In this study, reduction of genetic diversity during sorghum domestication was also observed in the C_4 gene families, indicating that wild sorghum can be explored for improving C_4 photosynthesis as a repository for genetic diversity.

However, the overall reduction in diversity of C_4 gene families was not significantly different from the genome-wide average, indicating that this gene family has not been under particularly strong selection pressure. Similarly, none of the 9 core C_4 genes showed a domestication signal at the gene level. The absence of large sequence variation at the gene level is also consistent with previous evolutionary studies suggesting that relatively minor changes to pre-existing regulatory networks and the use of pre-existing *cis*-elements were often sufficient to recruit genes into the C_4 pathway [43–45]. The C_4 isoform of the NADP-ME gene found in maize and sorghum is one such gene that has been found to be activated for C_4 photosynthesis via subtle changes to its promoter, while the rest of the gene is highly conserved [32]. This is consistent with the low diversity in this gene family observed in our study.

A further high-resolution investigation of domestication signature at the SNP level revealed 2 C_4 genes, PPDK (*Sobic.009G132900)* and CA (*Sobic.003G234200)* with SNPs under purifying selection, while the other 2 C_4 genes, PPDK-RP (*Sobic.002G324400)* and PEPC (*Sobic.010G160700)*,, were identified with SNPs under balancing selection. Previous studies have demonstrated that SNP-level analysis using less stringent criteria is superior for capturing soft selection signals compared with genome-wide ranking [42, 46]. However, the higher sensitivity may come with a cost of a greater chance of false positives, and therefore requires cautious interpretation. The contrasting selection signals on genes from the same pathway within taxa found in this study was also reported previously in signal transduction pathways [47] and the starch biosynthesis pathway [48].

The C_4 isoforms of PPDK and PEPC were also found to show signals of positive selection in a previous cross-species evolutionary study using orthologous groups from closely related C_3 and C_4 grassspecies including sorghum [27]. PPDK and PPDK-RP regulate the regeneration of PEP and as such have a direct effect on CO_2 assimilation rate [49], especially under cool temperatures [50, 51]. However, it is thought that only minor changes to the enzyme properties of PPDK were sufficient to recruit it into the C_4 pathway and its residues and regions involved in catalysis are highly conserved in C_4 species [52], possibly validating the fact that only soft selection signals via SNP-level were found for the C_4 isoform of the PPDK gene in our study.

PEPC is also regarded as a potential limiting step in assimilation of CO_2, and variation of its affinity for CO_2/HCO_3^- amongst species has been documented [53–55]. CA is also critical to C_4 photosynthesis as it catalyses the first step of the C_4 pathway, converting CO_2 to HCO_3^- [56]. It was reported in the C_4 dicot *Flaveria bidentis,* where antisense plants with <10% of wild-type CA activity required high CO_2 for growth and showed reduced CO_2 assimilation rates [57, 58]. Recent experiments showed CA and PEPC will be more limiting when stomata are partially closed, e.g. under water limitation [59].

The signal of soft purifying selection on PPDK and CA may suggest the C_4 pathway was indirectly improved during sorghum domestication. Without photosynthetic rate being a direct selection target in breeding programs, a steady increase in leaf photosynthetic rate over time of cultivar release has been shown in other cereals, e.g. in Australian bread wheat [60]. The balancing selection signal on C_4 PPDK-RP and PEPC may reflect adaptation to diverse environments, as both PPDK-RP and PEPC are associated with abiotic stress [61, 62]. Interestingly, within the PPDK-RP and PPDK gene families, the non-C_4 genes all showed selection signals contrasting with their C_4 counterparts with both two non-C_4 PPDK-RP (*Sobic.002G324500, Sobic.002G324700)* containing SNPs under purifying selection and the non-C_4 PPDK (*Sobic.001G326900)* containing SNPs under balancing selection.

After domestication, sorghum was introduced from tropical to temperate areas, and adapted to divergent local environments. New mutations also arose during this diversification process, and played an important role in local adaptation. In the haplotype analysis, these haplotypes unique to cultivated sorghum are likely to be young alleles arising after domestication, while haplotypes unique to the wild progenitor indicate that some haplotypes were lost during domestication of sorghum. Nevertheless, the loss of wild haplotypes of C_4 genes in cultivated sorghum does not mean these haplotypes are inferior in terms of photosynthetic efficiency, as photosynthesis was not specifically targeted during sorghum domestication [10]. On the contrary, bringing these wild haplotypes back to breeding programs after evaluation of their functions may enrich breeders' toolkits to manipulate photosynthetic efficiency, ultimately contributing to yield improvements.

C_4 photosynthesis has been well studied over the past 50 years and key components of this complex pathway have been identified following the advent of transgenic and sequencing technologies [8]. Understanding the genetic diversity of the key enzymes of the C_4 pathway is an important step towards mining the natural allelic variation for the improvement of photosynthesis. Further assessment of effects of different alleles in similar genetic backgrounds will move us closer to being able to exploit the natural allelic variation to improve crop yields.

Materials And Methods

Identification of C_4 gene families

A total of 9 key C_4 genes involved in the NADP-ME photosynthetic pathway (Figure 1) and their non-C_4 isoforms in sorghum were extracted from Williams et al.63] and Wang, et al. [64] (Table 1). Homology between these sorghum C_4 genes and their non-C_4 isoforms was further verified via a local blast strategy. Protein sequence of these 9 core C_4 genes were extracted from the sorghum reference genome V3.1 and were blasted against the reference genome. Blast hits of each gene were filtered using the criteria: E-value<1e–10, sequence identity >60% and alignment length>80%. All hits of the same gene satisfying the criteria were plotted based on –log (E-value), only hits of top –log (E-value) class were considered if clear differentiation among them was visualized, otherwise all hits were used.

Plant material and genomic data

Sequence data of the identified C_4 genes were extracted from 48 accessions of *Sorghum bicolor* with high mapping depth (~22X per accession, ranging from 16X to 45X) reported in previous studies [23–25]. These 48 accessions represent all major cultivated sorghum races and some wild progenitors (Table S1).

Gene- level population genetic analyses

Population genetic parameters including nucleotide diversity (θπ) [65], Tajima's D [66] and Watterson's Estimator (hW) [67] were directly calculated for each of the 27 genes using the Bio::PopGen::Statistics module. Fst [68] which measures population differentiation was also calculated for each of the 27 genes using the Bio::PopGen::PopStats moudel [25]. The Bio::PopGen::IO module was used to read the csv format input file, which was prepared using a in-house perl script for calculation of these population genetic parameters.

The criteria used in Mace et al. [25] were employed to identify genes under purifying selection and balancing selection, respectively. Criteria for purifying selection included: 1) θπ and hW <5% of the empirical distribution in the cultivated group, 2) Fst between the group of cultivated sorghum and the group of wild and weedy sorghum >95% of the population pairwise distribution, 3) Tajima's D <0. Criteria for balancing selection included: 1) θπ and hW >25% of the empirical distribution in the cultivated group, 2) Fst between the group of cultivated sorghum and the group of wild and weedy sorghum <90% of the population pairwise distribution, 3) Tajima's D > 5% of the empirical distribution.

SNP- level identification of selection signature

Population genetics parameters including θπ [37], Tajima's D [38] and Fst [40] between the group of cultivated sorghum and the group of wild and weedy sorghum were computed for these 27 genes using CDS in PopGenome, a population genomics package implemented in the R environment (http://cran.r-project.org/) [69]. Specifically, commands diversity.stats, F_ST.stats and neutrality.stats were called to calculate θπ, Fst and Tajima's D for each SNP, respectively, with a slide window of 1-bp and 1-bp step size. Functional annotation of each SNP was conducted using get.codons command. Fold decrease of θπ in the cultivated sorghum group compared to the group of wild and weedy sorghum was calculated to represent reduction of diversity (RoD). The following criteria were adopted to identify sites with signature of purifying selection: (1) A RoD greater than the average of neutral genes; (2) Fst > 0; (3) Tajima's D < 0. The following criteria were adopted to identify sites with signature of balancing selection: (1) an increase of diversity in the cultivated group and the group of wild and weedy comparison; (2) Fst > 0; (3) Tajima's D > 0.

Phylogenetic and Haplotype analysis.

Phylogenetic tree was constructed based on the CDS sequence of all 27 genes from the C_4 gene families using the neighbor-joining method in Geneious 8.1.2. Analysis of haplotype networks were conducted using a combination of the R package ape [70] and pegas [71]. All 48 sorghum accessions were classified into four groups: cultivated, wild and weedy, *Guinea margaritiferum* and *S. propinquum* (Table S2).

Abbreviations

CO_2: carbon dioxide

CDS coding sequence

RoD : reduction of diversity

Rubisco: Ribulose–1,5-bisphosphate carboxylase

SNP: single nucleotide polymorphism

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

All genetic variation data reported in this study can be retrieved from the sorghum SNP database, SorGSD (http://sorgsd.big.ac.cn/).

Competing interests

The authors declare that they have no competing interests

Acknowledgements

We thank Drs Susanne von Caemmerer and Robert Furbank for their valuable comments and suggestions towards the improvement of this manuscript.

Funding

This work was funded partially by the Australian Government through the Australian Research Council Centre of Excellence for Translational Photosynthesis (grant number

CE140100015) and State Key Laboratory of Agricultural Genomics, China (grant number 2011DQ782025).

Authors' contributions

D. J. and E. M. conceived the original idea. Y. T., M.B-P., S. T., A. C., and E. M. analysed the data. Y. T. and B.G-J. wrote the manuscript. All authors discussed the results and contributed to the final manuscript.

References

1.R. F. Sage, T. L. Sage, F. Kocacinar, *Photorespiration and the evolution of C_4 photosynthesis. Annu. Rev. Plant Biol.,* 63(2012), pp. 19–47.

2.R. F. Sage, *The evolution of C_4 photosynthesis. New Phytol.,* 161(2004), pp. 341–370.

3.E. J. Edwards, S. A. Smith, *Phylogenetic analyses reveal the shady history of C_4 grasses. Proc. Natl. Acad. Sci. USA,* 107(2010), pp. 2532–2537.

4.J. M. Hibberd, S. Covshoff, *The regulation of gene expression required for C_4 photosynthesis. Annu. Rev. Plant Biol.,* 61(2010), pp. 181–207.

5.S. von Caemmerer, R. T. Furbank, *The C_4 pathway: an efficient CO_2 pump. Photosynth. Res.,* 77(2003), pp. 191.

6.J. M. Hibberd, J. E. Sheehy, J. A. Langdale, *Using C_4 photosynthesis to increase the yield of rice-rationale and feasibility. Curr. Opin. Plant Biol.,* 11(2008), pp. 228–231.

7.J. A. Langdale, *C_4 cycles: past, present, and future research on C_4 photosynthesis. Plant Cell,* 23(2011), pp. 3879–3892.

8.S. von Caemmerer, R. T. Furbank, *Strategies for improving C_4 photosynthesis. Curr. Opin. Plant Biol.,* 31(2016), pp. 125–134.

9.C. J. Still, J. A. Berry, G. J. Collatz, R. S. DeFries, *Global distribution of C_3 and C_4 vegetation: carbon cycle implications. Global Biogeochem. Cy.,* 17(2003), pp. 1–14.

10.S. P. Long, X. G. Zhu, S. L. Naidu, D. R. Ort, *Can improvement in photosynthesis increase crop yields? Plant Cell Environ.,* 29(2006), pp. 315–330.

11.X. G. Zhu, L. L. Shan, Y. Wang, W. P. Quick, *C_4 rice-an ideal arena for systems biology research. J. Integr. Plant Biol.,* 52(2010), pp. 762–770.

12.S. von Caemmerer, W. P. Quick, R. T. Furbank, *The development of C_4 rice: current progress and future challenges. Science,* 336(2012), pp. 1671–1672.

13.S. Covshoff, M. Szecowka, T. E. Hughes, R. Smith-Unna, S. Kelly, K. J. Bailey, T. L. Sage, J. A. Pachebat, R. Leegood, J. M. Hibberd, *C_4 photosynthesis in the rice paddy: insights from the noxious weed echinochloa glabrescens. Plant Physiol.,* 170(2016), pp. 57–73.

14.M. D. Hatch, T. Kagawa, S. Craig, *Subdivision of C_4-pathway species based on differing C_4 acid decarboxylating systems and ultrastructural features. Funct. Plant Biol.,* 2(1975), pp. 111–128.

15.A. Bräutigam, S. Schliesky, C. Külahoglu, C. P. Osborne, A. P.M. Weber, *Towards an integrative model of C_4 photosynthetic subtypes: insights from comparative transcriptome analysis of NAD-ME, NADP-ME, and PEP-CK C_4 species. J. Exp. Bot.,* 65(2014), pp. 3579–3593.

16.B. V. Sonawane, R. E. Sharwood, S. Whitney, O. Ghannoum, *Shade compromises the photosynthetic efficiency of NADP-ME less than PEP-CK and NAD-ME C_4 grasses. J. Exp. Bot.* 69(2018), pp. 3053–3068.

17.G. Grass Phylogeny Working, II, *New grass phylogeny resolves deep evolutionary relationships and discovers C_4 origins. New Phytol.,* 193(2012), pp. 304–312.

18.S. P. Kidambi, D. R. Krieg, D. T. Rosenow, *Genetic variation for gas exchange rates in grain sorghum. Plant Physiol.,* 92(1990), pp. 1211–1214.

19.S. B. Peng, D. R. Krieg, *Gas exchange traits and their relationship to water-use efficiency of grain sorghum. Crop Sci.,* 32(1992), pp. 386–391.

20.S. Henderson, S. von Caemmerer, G. D. Farquhar, L. J. Wade, G. Hammer, *Correlation between carbon isotope discrimination and transpiration efficiency in lines of the C_4 species Sorghum bicolor in the glasshouse and the field. Aust. J. Plant Physiol.,* 25(1998), pp. 111–123.

21.M. Balota, W. A. Payne, W. Rooney, D. Rosenow, *Gas exchange and transpiration ratio in sorghum. Crop Sci.,* 48(2008), pp. 2361–2371.

22.M. G. S. Fernandez, K. Strand, M. T. Hamblin, M. Westgate, E. Heaton, S. Kresovich, *Genetic analysis and phenotypic characterization of leaf photosynthetic capacity in a sorghum (Sorghum spp.) diversity panel. Genet. Resour. Crop Ev.,* 62(2015), pp. 939–950.

23.L. Y. Zheng, X. S. Guo, B. He, L. J. Sun, Y. Peng, S. S. Dong, T. F. Liu, S. Y. Jiang, S. Ramachandran, C. M. Liu, H. C. Jing, *Genome-wide patterns of genetic variation in sweet and grain sorghum (Sorghum bicolor). Genome Biol.,* 12(2011), pp. 1–14.

24.A. H. Paterson, J. E. Bowers, R. Bruggmann, I. Dubchak, J. Grimwood, H. Gundlach, G. Haberer, U. Hellsten, T. Mitros, A. Poliakov, J. Schmutz, M. Spannagl, H. B. Tang, X. Y. Wang, T. Wicker, A. K. Bharti, J. Chapman, F. A. Feltus, U. Gowik, I. V. Grigoriev, E. Lyons, C. A. Maher, M. Martis, A. Narechania, R. P. Otillar, B. W. Penning, A. A. Salamov, Y. Wang, L. F. Zhang, N. C. Carpita, M. Freeling, A. R. Gingle, C. T. Hash, B. Keller, P. Klein, S. Kresovich, M. C.

McCann, R. Ming, D. G. Peterson, Mehboob-ur-Rahman, D. Ware, P. Westhoff, K. F. X. Mayer, J. Messing, D. S. Rokhsar, *The Sorghum bicolor genome and the diversification of grasses. Nature,* 457(2009), pp. 551–556.

25.E. S. Mace, S. S. Tai, E. K. Gilding, Y. H. Li, P. J. Prentis, L. L. Bian, B.C. Campbell, W. S. Hu, D. J. Innes, X. L. Han, A. Cruickshank, C. M. Dai, C. Frere, H. K. Zhang, C. H. Hunt, X. Y. Wang, T. Shatte, M. Wang, Z. Su, J. Li, X. Z. Lin, I. D. Godwin, D. R. Jordan, J. Wang, *Whole-genome sequencing reveals untapped genetic potential in Africa's indigenous cereal crop sorghum. Nat. Commun.,* 4(2013), pp. 1–9.

26.N. Fankhauser, S. Aubry, *Post-transcriptional regulation of photosynthetic genes is a key driver of C_4 leaf ontogeny. J. Exp. Bot.,* 68(2017), pp. 137–146.

27.P. Huang, A. J. Studer, J. C. Schnable, E. A. Kellogg, T. P. Brutnell, *Cross species selection scans identify components of C_4 photosynthesis in the grasses. J. Exp. Bot.,* 68(2017), pp. 127–135.

28.S. J. Burgess, J. M. Hibberd, *Insights into C_4 metabolism from comparative deep sequencing. Curr. Opin. Plant Biol.,* 25(2015), pp. 138–144.

29.G. Reeves, M. J. Grangé-Guermente, J. M. Hibberd, *Regulatory gateways for cell-specific gene expression in C_4 leaves with Kranz anatomy. J. Exp. Bot.,* 68(2017), pp. 107–116.

30.P.-A. Christin, C. P. Osborne, D. S. Chatelet, J. T. Columbus, G. Besnard, T. R. Hodkinson, L. M. Garrison, M. S. Vorontsova, E. J. Edwards, *Anatomical enablers and the evolution of C_4 photosynthesis in grasses. Proc. Natl. Acad. Sci. USA,* 110(2013), pp. 1381–1386.

31.J. J. Moreno-Villena, L. T. Dunning, C. P. Osborne, P.-A. Christin, *Highly expressed genes are preferentially co-opted for C_4 photosynthesis. Mol. Biol. Evol.,* 35(2018), pp. 94–106.

32.A. R. Borba, T. S. Serra, A. Gorska, P. Gouveia, A.M. Cordeiro, I. Reyna-Llorens, J. Knerova, P.M. Barros, I. A. Abreu, M. M. O. Oliveira, J. M. Hibberd, N. J. M. Saibo, *Synergistic binding of bHLH transcription factors to the promoter of the maize NADP-ME gene used in C_4 photosynthesis is based on an ancient code found in the ancestral C_3 state. Mol. Biol. Evol.* 7(2018), pp. 1690–1705.

33.P. A. Christin, B. Petitpierre, N. Salamin, L. Büchi, G. Besnard, *Evolution of C_4 phospho enol pyruvate carboxykinase in grasses, from genotype to phenotype. Mol. Biol. Evol.,* 26(2008), pp. 357–365.

34.P. A. Christin, N. Salamin, A.M. Muasya, E. H. Roalson, F. Russier, G. Besnard, *Evolutionary switch and genetic convergence on rbcL following the evolution of C_4 photosynthesis. Mol. Biol. Evol.,* 25(2008), pp. 2361–2368.

35.J. R. Ehleringer, R. F. Sage, L. B. Flanagan, R. W. Pearcy, *Climate change and the evolution of C_4 photosynthesis. Trends Ecol. Evol.,* 6(1991), pp. 95–99.

36.P. A. Christin, N. Salamin, V. Savolainen, M. R. Duvall, G. Besnard, *C_4 photosynthesis evolved in grasses via parallel adaptive genetic changes. Curr. Biol.,* 17(2007), pp. 1241–1247.

37.P. A. Christin, E. Samaritani, B. Petitpierre, N. Salamin, G. Besnard, *Evolutionary insights on C_4 photosynthetic subtypes in grasses from genomics and phylogenetics. Genome Biol. Evol.,* 1(2009), pp. 221–230.

38.J. D. Clark, A. Stemler, *Early domesticated sorghum from central Sudan. Nature,* 254(1975), pp. 588–591.

39.F. Wendorf, A. E. Close, R. Schild, K. Wasylikowa, R. A. Housley, J. R. Harlan, H. Królik, *Saharan exploitation of plants 8,000 years BP. Nature,* 359(1992), pp. 721–724.

40.Y. Tao, E. Mace, B. George-Jaeggli, C. Hunt, A. Cruickshank, R. Henzell, D. Jordan, *Novel grain weight loci revealed in a cross between cultivated and wild sorghum. Plant Genome,* 11(2018), pp. 1–10.

41.Y. Tao, E. S. Mace, S. Tai, A. Cruickshank, B.C. Campbell, X. Zhao, E. J. Van Oosterom, I. D. Godwin, J. R. Botella, D. R. Jordan, *Whole-genome analysis of candidate genes associated with seed size and weight in Sorghum bicolor reveals signatures of artificial selection and insights into parallel domestication in cereal crops. Front. Plant Sci.,* 8(2017), pp. 1–14.

42.Y. Tao, X. Zhao, E. Mace, R. Henry, D. Jordan, *Exploring and exploiting pan-genomics for crop improvement. Mol. Plant,* 12(2019), pp. 156–169.

43.B. M. C. Kümpers, S. J. Burgess, I. Reyna-Llorens, R. Smith-Unna, C. Boursnell, J. M. Hibberd, *Shared characteristics underpinning C_4 leaf maturation derived from analysis of multiple C_3 and C_4 species of Flaveria. J. Exp. Bot.,* 68(2017), pp. 177–189.

44.C. Külahoglu, A. K. Denton, M. Sommer, J. Maß, S. Schliesky, T. J. Wrobel, B. Berckmans, E. Gongora-Castillo, C. R. Buell, R. Simon, L. De Veylder, A. Bräutigam, A. P.M. Weber, *Comparative transcriptome atlases reveal altered gene expression modules between two cleomaceae C_3 and C_4 plant species. Plant Cell,* 26(2014), pp. 3243–3260.

45.S. J. Burgess, I. Reyna-Llorens, S. R. Stevenson, P. Singh, K. Jaeger, J. M. Hibberd, *Genome-wide transcription factor binding in leaves from C_3 and C_4 grasses. bioRxiv (*2019), pp. 165787.

46.K. Massel, B.C. Campbell, E. S. Mace, S. Tai, Y. Tao, B. G. Worland, D. R. Jordan, J. R. Botella, I. D. Godwin, *Whole genome sequencing reveals potential new targets for improving nitrogen uptake and utilization in Sorghum bicolor. Front. Plant Sci.,* 7(2016), pp. 1544.

47.R. M. Riley, W. Jin, G. Gibson, *Contrasting selection pressures on components of the Ras-mediated signal transduction pathway in Drosophila. Mol. Ecol.,* 12(2003), pp. 1315–1323.

48.B.C. Campbell, E. K. Gilding, E. S. Mace, S. Tai, Y. Tao, P. J. Prentis, P. Thomelin, D. R. Jordan, I. D. Godwin, *Domestication and the storage starch biosynthesis pathway: Signatures of selection from a whole sorghum genome sequencing strategy. Plant Biotechnol. J.* 14(2016), pp.2240–2253.

49.Y. Wang, W. Xu, L. Hu, L. Zhang, Y. Li, X. Du, *Expression of maize gene encoding C_4-pyruvate orthophosphate dikinase (PPDK) and C_4-phosphoenolpyruvate carboxylase (PEPC) in transgenic Arabidopsis. Plant Mol. Biol. Rep.,* 30(2012), pp. 1367–1374.

50.D. Wang, A. R. Portis, S. P. Moose, S. P. Long, *Cool C_4 photosynthesis: pyruvate pi dikinase expression and activity corresponds to the exceptional cold tolerance of carbon assimilation in Miscanthus × giganteus. Plant Physiol.,* 148(2008), pp. 557–567.

51.S. L. Naidu, S. P. Moose, A. K. AL-Shoaibi, C. A. Raines, S. P. Long, *Cold tolerance of C_4 photosynthesis in Miscanthus × giganteus: adaptation in amounts and sequence of C_4 photosynthetic enzymes. Plant Physiol.,* 132(2003), pp. 1688–1697.

52.C. J. Chastain, C. J. Failing, L. Manandhar, M. A. Zimmerman, M. M. Lakner, T. H. T. Nguyen, *Functional evolution of C_4 pyruvate, orthophosphate dikinase. J. Exp. Bot.,* 62(2011), pp. 3083–3091.

53.H. Bauwe, R. Chollet, *Kinetic properties of phosphoenolpyruvate carboxylase from C_3, C_4, and C_3-C_4 intermediate species of Flaveria (Asteraceae). Plant Physiol.,* 82(1986), pp. 695–699.

54.S. von Caemmerer, G. E. Edwards, N. Koteyeva, A. B. Cousins, *Single cell C_4 photosynthesis in aquatic and terrestrial plants: a gas exchange perspective. Aquat. Bot.,* 118(2014), pp. 71–80.

55.R. A. Boyd, A. Gandin, A. B. Cousins, *Temperature responses of C_4 photosynthesis: biochemical analysis of rubisco, phosphoenolpyruvate carboxylase, and carbonic anhydrase in Setaria viridis. Plant Physiol.,* 169(2015), pp. 1850–1861.

56.M. D. Hatch, J. N. Burnell, *Carbonic anhydrase activity in leaves and its role in the first step of C_4 photosynthesis. Plant Physiol.,* 93(1990), pp. 825–828.

57.S. Von Caemmerer, V. Quinn, N. Hancock, G. D. Price, R. T. Furbank, M. Ludwig, *Carbonic anhydrase and C_4 photosynthesis: a transgenic analysis. Plant Cell Environ.,* 27(2004), pp. 697–703.

58.A. B. Cousins, M. R. Badger, S. von Caemmerer, *Carbonic anhydrase and its influence on carbon isotope discrimination during C_4 photosynthesis. Insights from antisense RNA in Flaveria bidentis. Plant Physiol.,* 141(2006), pp. 232–242.

59.H. L. Osborn, H. Alonso-Cantabrana, R. E. Sharwood, S. Covshoff, J. R. Evans, R. T. Furbank, S. von Caemmerer, *Effects of reduced carbonic anhydrase activity on CO_2 assimilation rates in Setaria viridis: a transgenic analysis. J. Exp. Bot.,* 68(2016), pp. 299–310.

60.N. Watanabe, J. R. Evans, W. S. Chow, *Changes in the photosynthetic properties of Australian wheat cultivars over the last century. Aust. J. Plant Physiol.,* 21(1994), pp. 169–183.

61.X. Liu, X. Li, C. Zhang, C. Dai, J. Zhou, C. Ren, J. Zhang, *Phosphoenolpyruvate carboxylase regulation in C_4-PEPC-expressing transgenic rice during early responses to drought stress. Physiol. plant.,* 159(2017), pp. 178–200.

62.M. Jeanneau, D. Gerentes, X. Foueillassar, M. Zivy, J. Vidal, A. Toppan, P. Perez, *Improvement of drought tolerance in maize: towards the functional validation of the Zm-Asr1 gene and increase of water use efficiency by over-expressing C_4–PEPC. Biochimie,* 84(2002), pp. 1127–1135.

63.B. P. Williams, S. Aubry, J. M. Hibberd, *Molecular evolution of genes recruited into C_4 photosynthesis. Trends Plant Sci.,* 17(2012), pp. 213–220.

64.X. Wang, U. Gowik, H. Tang, J. E. Bowers, P. Westhoff, A. H. Paterson, *Comparative genomic analysis of C_4 photosynthetic pathway evolution in grasses. Genome Biol., 10(*2009), pp. R68.

65.M. Nei, W. Li, *Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc. Natl. Acad. Sci. USA,* 76(1979), pp. 5269–5273.

66.F. Tajima, *Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics,* 123(1989), pp. 585–595.

67.G. A. Watterson, *On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol.,* 7(1975), pp. 256–276.

68.R. Hudson, D. D. Boos, N. Kaplan, *A statistical test for detecting geographic subdivision. Mol. Biol. Evol.,* 9(1992), pp. 138–151.

69.B. Pfeifer, U. Wittelsbürger, S. E. R. Onsins, M. J. Lercher, *PopGenome: an efficient Swiss army knife for population genomic analyses in R. Mol. Biol. Evol. (*2014), pp. msu136.

70.E. Paradis, J. Claude, K. Strimmer, *APE: analyses of phylogenetics and evolution in R language. Bioinformatics,* 20(2004), pp. 289–290.

71.E. Paradis, *pegas: an R package for population genetics with an integrated-modular approach. Bioinformatics,* 26(2010), pp. 419–420.

Tables

Table 1 Single nucleotide polymorphism (SNP) information and selection signals across 27 genes from C_4 gene families

Gene ID	Enzyme	GL	CDSL	NoS	NoSiC	NoNS	NoSS	UPSGL	UBSGL	NoSUPS	NoNSUPS	NoSUBS	NoNSUBS
Sobic.002G230100	CA	4823	1014	115	14	4	10	No	No	0	0	1	0
Sobic.003G234200	CA	10440	1371	475	33	7	26	No	No	1	1	0	0
Sobic.003G234400	CA	4749	615	138	13	3	10	No	No	0	0	0	0
Sobic.003G234500	CA	2986	609	173	11	5	6	No	No	0	0	0	0
Sobic.003G234600	CA	4750	771	210	18	10	8	No	No	0	0	0	0
Sobic.007G166200	NADP-MDH	3354	1308	53	11	6	5	No	No	0	0	0	0
Sobic.007G166300	NADP-MDH	3816	1290	108	12	3	9	No	No	0	0	0	0
Sobic.003G036000	NADP-ME	6107	1941	111	11	4	7	No	No	0	0	0	0
Sobic.003G036200	NADP-ME	5447	1911	141	12	3	9	No	No	0	0	0	0
Sobic.003G280900	NADP-ME	5691	1782	175	22	13	9	No	No	1	1	0	0
Sobic.003G292400	NADP-ME	4527	1782	95	22	8	14	No	No	10	2	0	0
Sobic.009G069600	NADP-ME	3624	1713	118	34	10	24	No	No	3	1	0	0
Sobic.002G167000	PEPC	5632	2904	41	11	6	5	No	No	0	0	0	0
Sobic.003G100600	PEPC	8881	3117	371	43	9	34	No	No	0	0	21	2
Sobic.003G301800	PEPC	7610	2901	138	19	3	17	No	No	0	0	0	0
Sobic.004G106900	PEPC	6977	2883	146	34	5	29	No	No	0	0	7	0
Sobic.007G106500	PEPC	5616	2895	64	12	8	4	No	No	1	1	0	0
Sobic.010G160700	PEPC	6647	3087	193	28	9	19	No	No	0	0	2	0
Sobic.004G219900	PPCK	1612	924	40	9	1	8	No	No	0	0	2	0
Sobic.004G338000	PPCK	1749	855	37	9	4	4	No	No	0	0	0	0
Sobic.006G148300	PPCK	1997	900	64	4	1	3	No	No	0	0	0	0
Sobic.001G326900	PPDK	8494	2730	321	46	18	28	No	Yes	0	0	24	5
Sobic.009G132900	PPDK	12748	2847	441	16	0	16	No	No	3	0	0	0
Sobic.002G324400	PPDK-RP	2507	1290	79	22	8	14	No	No	0	0	3	0
Sobic.002G324500	PPDK-RP	3072	1260	69	20	5	15	No	No	4	0	0	0
Sobic.002G324700	PPDK-RP	4662	1587	222	28	19	9	No	No	1	1	2	2
Sobic.005G042000	RbcS	1556	510	45	7	4	3	No	No	0	0	0	0

Gene ID is according to sorghum reference genome V3.1. Gene IDs in bold indicate their C_4 genes. Enzyme: encoded enzyme. GL: gene length. CDSL: length of CDS. NoS: total number of SNPs identified across the gene. NoSiC: number of SNPs identified in CDS. NoNS: number of non-synonymous SNPs. NoSS: number of synonymous SNPs. UPSGL: under purifying selection based on gene level analysis. UBSGL: under balancing selection based on gene level analysis. NoSUPS:

number of SNPs under purifying selection. NoNSUPS: number of non-synonymous SNPs under purifying selection. NoSUBS: number of SNPs under balancing selection. NoNSUBS: number of non-synonymous SNPs under balancing selection.

Table 2 Genetic diversity (θπ) and fixation index (Fst) of 27 genes from C_4 gene families

GeneID	Enzyme	θπ-All	θπ-Cultivated	θπ-W&W	Fst
Sobic.002G230100	CA	0.80	0.74	0.90	0.19
Sobic.003G234200	CA	2.65	2.46	2.66	0.16
Sobic.003G234400	CA	1.01	0.91	0.88	0.37
Sobic.003G234500	CA	5.55	5.51	4.56	0.07
Sobic.003G234600	CA	1.27	1.35	0.65	0.06
Sobic.007G166200	NADP-MDH	0.18	0.21	0.13	0.07
Sobic.007G166300	NADP-MDH	0.33	0.33	0.42	0.08
Sobic.003G036000	NADP-ME	0.88	0.65	1.59	0.15
Sobic.003G036200	NADP-ME	0.89	0.67	1.39	0.06
Sobic.003G280900	NADP-ME	0.93	0.85	1.11	0.09
Sobic.003G292400	NADP-ME	1.43	0.08	4.44	0.32
Sobic.009G069600	NADP-ME	0.52	0.49	0.10	0.45
Sobic.002G167000	PEPC	0.58	0.51	0.85	0.04
Sobic.003G100600	PEPC	5.36	5.18	3.56	0.05
Sobic.003G301800	PEPC	0.64	0.22	2.37	0.22
Sobic.004G106900	PEPC	3.18	3.02	2.14	0.07
Sobic.007G106500	PEPC	0.44	0.22	0.47	0.21
Sobic.010G160700	PEPC	2.49	2.25	2.86	0.04
Sobic.004G219900	PPCK	2.08	1.94	2.12	0.12
Sobic.004G338000	PPCK	1.03	0.96	0.91	0.03
Sobic.006G148300	PPCK	0.48	0.39	0.13	0.41
Sobic.001G326900	PPDK	8.34	5.64	5.64	0.40
Sobic.009G132900	PPDK	2.07	1.79	2.19	0.13
Sobic.002G324400	PPDK-RP	5.04	3.82	4.55	0.41
Sobic.002G324500	PPDK-RP	1.27	0.10	3.75	0.24
Sobic.002G324700	PPDK-RP	2.58	2.50	3.51	0.05
Sobic.005G042000	rbcS	4.32	3.41	5.72	0.12

Gene ID is according to sorghum reference genome V3.1. Gene IDs in bold indicate the C_4 gene versions. Enzyme: encoded enzyme. θπ-All: nucleotide diversity across all 48 genotypes. θπ-Cultivated: nucleotide diversity across cultivated genotypes. θπ-W&W: nucleotide diversity across wild & weedy genotypes. All θπ values are in unites of per kb. Fst: fixation index between cultivated genotypes and wild & weedy genotypes.

Figures

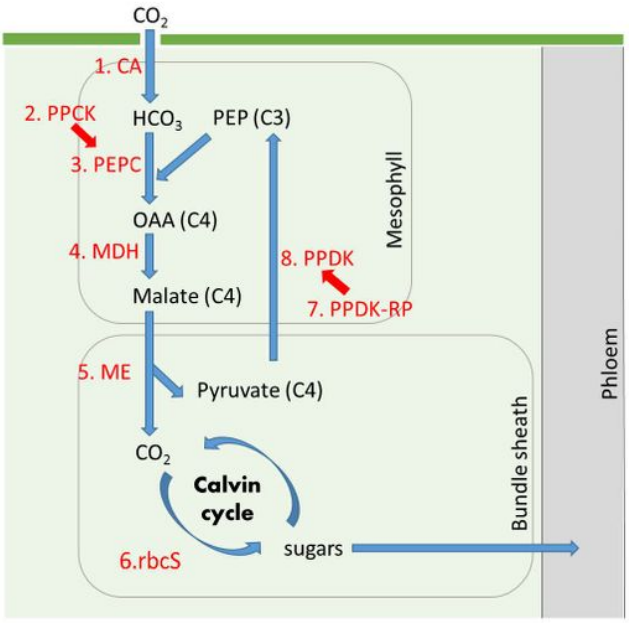

Figure 1

Diagram of the NADP-ME biosynthetic pathway of C4 photosynthesis (adapted from Wang 2009). In the mesophyll cells, CO2 is converted to HCO3- catalysed by Carbonic anhydrase (CA) and fixed into the four-carbon acid, oxaloacetate (OAA), by phosphoenolpyruvate carboxylase (PEPC). The OAA generated by PEPC is then reduced to malate by the NADP-malate dehydrogenase (NADP-MDH) or trans-aminated to aspartate. The resultant C4 acids, malate and aspartate, are transported to the bundle sheath and then decarboxylated in the vicinity of Rubisco to release CO2 and pyruvate. Pyruvate is transported back to mesophyll cells to regenerate PEP by pyruvate orthophosphate dikinase (PPDK), while CO2 enters the Calvin-Benson-Bassham cycle and is fixed by Rubisco.

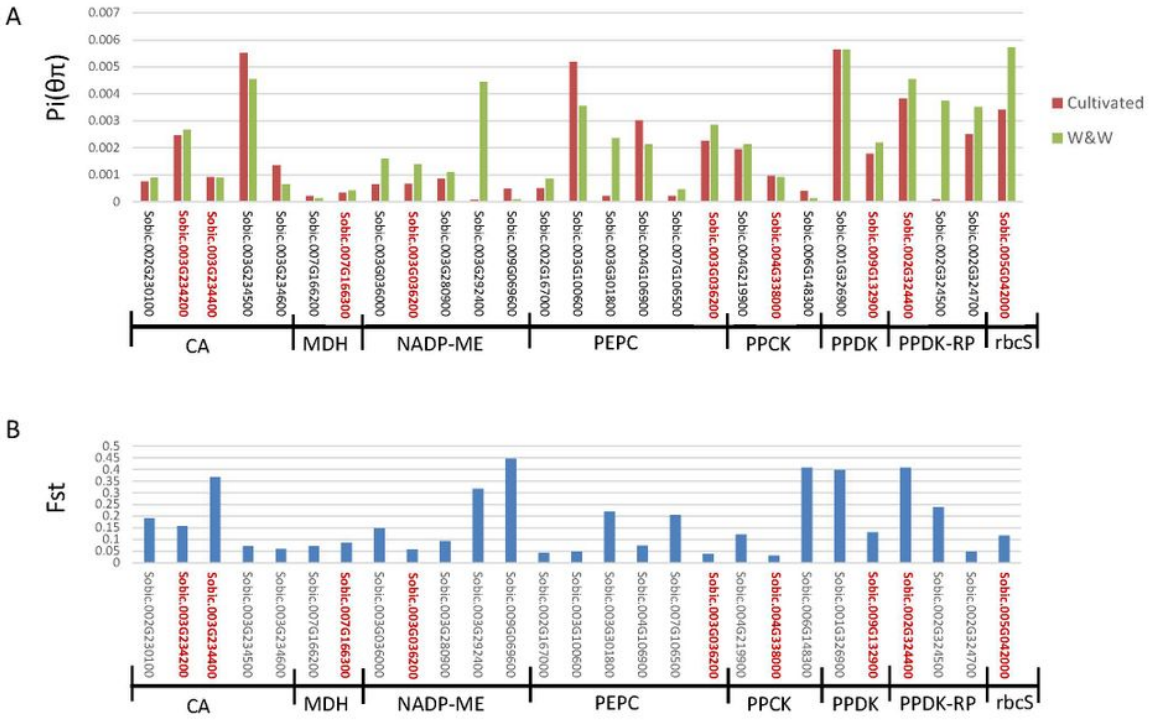

Figure 2

Genetic diversity and fixation index (Fst) of C4 gene families between cultivated sorghum and the wild and weedy group. (A) genetic diversity (pi) for each of the C4 gene families. Gene IDs in red indicate core C4 genes. Red bars represent Pi of cultivated sorghum, while green bars represent Pi of wild and weedy. (B), Fst between cultivated and wild and weedy of each of C4 gene families. Gene IDs in red indicate core C4 genes.

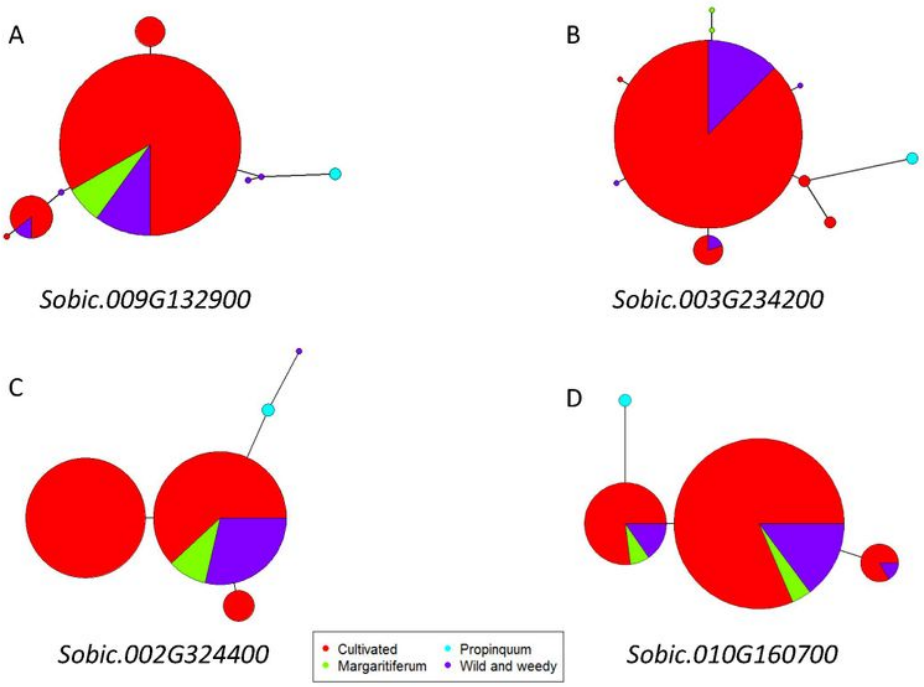

Figure 3

Haplotype network of 4 core C4 gene with selection signal based on individual SNP analysis. (A) the PPDK gene (Sobic.009G132900) with signal of purifying selection; (B) one of the CA genes (Sobic.003G234200) with signal of purifying selection; (C) The PPDK-RP gene (Sobic.002G324400) with signal of balancing selection; (D) the PEPC gene (Sobic.010G160700) with signal of balancing selection. Group classification of sorghum accessions used as detailed in Table S1. Colour-coding as follows; cultivated sorghum (red), wild and weedy genotypes (purple), Sorghum propinquum (blue), and Sorghum guinea margaritiferum (green). The size of the circles in the haplotype networks is proportionate to the number of accessions with that haplotype. The branch length represents the genetic distance between two haplotypes.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupTableS4.xlsx
- SupTableS3.xlsx
- SuppFig1Ed.pdf
- SupTableS2.xlsx
- SupTableS1.xlsx
- SupTableS5.xlsx