8-2019

An Integrated Genomics And Phenomics Approach To Study The Evolution Of C4 Photosynthesis

Daniel Santana de Carvalho
University of Nebraska-Lincoln

Follow this and additional works at: https://digitalcommons.unl.edu/agronhortdiss

Part of the Agricultural Science Commons, Agriculture Commons, Agronomy and Crop Sciences Commons, Botany Commons, Horticulture Commons, Other Plant Sciences Commons, and the Plant Biology Commons

https://digitalcommons.unl.edu/agronhortdiss/170

This Article is brought to you for free and open access by the Agronomy and Horticulture Department at DigitalCommons@University of Nebraska - Lincoln. It has been accepted for inclusion in Theses, Dissertations, and Student Research in Agronomy and Horticulture by an authorized administrator of DigitalCommons@University of Nebraska - Lincoln.
AN INTEGRATED GENOMICS AND PHENOMICS APPROACH TO STUDY THE EVOLUTION OF C₄ PHOTOSYNTHESIS

by

Daniel Santana de Carvalho

A DISSERTATION

Presented to the Faculty of
The Graduate College at the University of Nebraska
In Partial Fulfilment of Requirements
For the Degree of Doctor of Philosophy

Major: Agronomy and Horticulture

Under the Supervision of Professor James C. Schnable

Lincoln, Nebraska

August, 2019
AN INTEGRATED GENOMICS AND PHENOMICS APPROACH TO
STUDY THE EVOLUTION OF C₄ PHOTOSYNTHEIS

Daniel Santana de Carvalho, Ph.D.
University of Nebraska, 2019

Adviser: James C. Schnable

The C₄ photosynthetic pathway was first described over 50 years ago. Today, it is known that C₄ evolved independently > 60 in plant lineages, which involves understanding not only the genetic, but also the metabolic features and differences involved in this process. Also, several adaptations are involved in the evolution of this type of photosynthesis, for example: changes in leaf anatomy and the evolution of kranz anatomy, physiology and metabolic pathways. In order to further investigate this pathway, different technologies and methods have been developed to unravel genes involved in C₄ photosynthesis. With the advances in molecular biology and bioinformatics tools new approaches have been applied to increase the knowledge of different features of C₄ photosynthesis, such as: gene duplications, neofunctionalization, gene regulation and metabolic reactions. One of the biggest areas of interest concerning C₄ photosynthesis is improving crop yield by engineering the C₄ pathway into C₃ crops, such as rice and wheat. Here I present different approaches that can be used to study C₄ photosynthesis from both phenomics and genomics perspectives.
ACKNOWLEDGMENTS

I want to thank my advisor Dr. James Schnable for all the support throughout my PhD program and for the insightful science discussions. Thanks to your guidance I learned very important skills as a scientist, teacher and advisor. I also want to thank my graduate supervisory committee: Drs. Jennifer Clarke, Jeff Mower and Josh Herr for the support and guidance during my program.

It is also important to acknowledge my former advisors and collaborators who contributed to my academic career (in no particular order): Drs. Ana Almeida, Charbel El-Hani, Gilberto Bomfim, Aristoteles Goes-Neto, Roberto Andrade, Suani Pinho, Thierry Lobao, Angelo Duarte and Sunil Kumar. Your input and insights contributed greatly to my scientific career development.

My appreciation also extends to the Department of Agronomy and Horticulture and the UNL Plant Science Initiative (UNL-PSI) for providing me with the infrastructure needed to develop my research and for giving me the opportunity to work and study in such a great institution. My academic success has a lot to do with the support from amazing staff members. Lisa Vonfeldt, Danielle Lopez and Marlene Busse, thank you for helping and providing me with guidance on the non-scientific matters. Hats off to the Schnable Lab members for all the support and help with my projects. Every suggestion was really important for both my PhD projects and scientific career.

Last, but not least, I want to thank my family and friends for the support and understanding my absence on birthdays and special events. Special thanks to my parents, Vera and Antonio Carvalho, and my sister, Andrea Catala, for giving me the strength to keep going and for supporting my dreams. I love you! Thank you to both my Brazilian and Lincoln friends, your support was extremely important throughout these years far from home, specially during tough times.
The results presented in Chapter 2, Sections 2.1-2.5 have been published in Evolutionary Bioinformatics. (D. S. Carvalho, J. C. Schnable and A. M. R. Almeida, Evolutionary Bioinformatics, (January 2018).)

The results presented in Chapter 3 have been submitted for publication in Plant Direct, and is currently published as a pre-print on BioRxiv. (D. S. Carvalho and J. C. Schnable, bioRxiv (2018).)

Some of the content discussed in Chapter 4 is expected to be included in a manuscript that is currently in preparation for publication under the lead of Daniel Carvalho. (D. S. Carvalho, S. K. K. Raju, Y. Zhang and J. C. Schnable, "Reconstructed protein sequence evolution suggests C₄ photosynthesis evolved via a C₂ ancestor in the Paniceae," (in preparation for submission to refereed journal).)
# Table of Contents

## List of Figures  

## List of Tables  

## 1: Literature Review  

1.1 Overview and importance of C₄ photosynthesis  

1.2 Parallel evolution and specific changes  

1.3 Approaches for comparative analysis  

1.3.1 Phylogenetics  

1.3.2 Gene by gene analysis  

1.3.3 Whole genome comparison  

1.3.4 Whole transcriptome comparison and gene regulation  

1.3.5 C₄ pathway modeling  

1.4 Engineering the C₄ pathway  

1.5 Concluding remarks  

## 2: Integrating phylogenetic and network approaches to study gene family evolution: the case of the AGAMOUS family of floral genes  

2.1 Introduction  

2.2 Methods  

2.3 Results  

2.4 Discussion
2.5 Conclusions ........................................ 33
2.6 Application on a C₄ photosynthesis gene family: PEPC genes .... 35
2.7 Supplementary information ................................ 38

3: Isoseq transcriptome assembly of new C₃ panicoid grasses provides insights about evolution of Poaceae .................................................. 45
   3.1 Introduction ........................................ 45
   3.2 Methods ........................................... 49
   3.3 Results and Discussion ............................. 54
   3.4 Data availability .................................... 60
   3.5 Supplemental figures ................................ 60

4: Reconstructed protein sequence evolution suggests C₄ photosynthesis evolved via a C₂ ancestor in the Paniceae ................................. 63
   4.1 Introduction ........................................ 63
   4.2 Material and methods ............................... 66
   4.3 Results ............................................. 71
   4.4 Discussion ......................................... 79

References ................................................. 88

Appendix .................................................. 115

Appendix A: High-throughput imaging and phenotyping dataset of C₄ grain crops 116
   A.1 Introduction ....................................... 116
   A.2 Methods .......................................... 118
   A.3 Data description and potential use .................... 121
   A.4 Limitations ........................................ 123
   A.5 Supplemental material ................................ 126
## List of Figures

1.1 Schematic representation of the three C₄ photosynthetic subtypes .......................... 5

2.1 Phylogenetic tree of the AGAMOUS family genes. ..................................................... 28

2.2 Summary of network findings ..................................................................................... 30

2.3 Schematic diagram of results based on phylogenetic (left) and network (right) analyses. ........................................................................................................................................... 34

2.4 Summary of network and phylogenetic results for PEPC gene. ................................. 37

2.S1 Networks obtained at 75% and 67% ...................................................................... 43

2.S2 Network obtained at 60% .......................................................................................... 44

3.1 Current literature consensus phylogeny of the relationships between the grass species studied. ................................................................................................................. 48

3.2 A GEvo panel showing transcript coverage of the C₄ PPDK gene in *S. bicolor* Sobic.009G132900 in each of the three species tested. ......................................................... 59

3.3 Seven hundred distinct phylogenetic trees calculated from separate multiple sequence alignments of 275 putatively orthologous gene groups with large regions of alignment scored as high quality. .......................................................... 60

3.S1 Distribution of lengths for polished transcript sequences ........................................ 61

3.S2 Distribution of alignment lengths after GBlocks cleaning ....................................... 62

3.S3 Plot of 10,876 phylogenetic trees .............................................................................. 62

4.1 Models explaining evolution of different subtypes of C₄ photosynthesis. ............. 67
4.2 Simplified pathway representation of the three main C₄ photosynthesis subtypes. ................................. 72
4.3 Models showing different types of selection and how they affect dN/dS ratios. ......................................... 73
4.4 Unrooted phylogenetic trees of C₄ photosynthesis core enzymes, present in all subtypes, according to citations in Table 4.1. ................................................. 74
4.5 Examples of C₄ photosynthesis subtype specific enzymes unrooted phylogenetic trees. Branch lengths are equal. .......................................................... 76
4.6 Photosynthetic parameters measurements for C₄ photosynthesis subtypes ....................................... 77
4.7 Photosynthetic parameters measurements for five different origins of C₄ photosynthesis ......................... 78
4.8 Simplified pathway representation of the three main C₄ photosynthesis subtypes including the C₃, C₂ and photorespiratory pathways. ........................................ 85
4.9 Unrooted phylogenetic trees of C₄ photosynthesis photorespiratory enzymes. ................................... 86
4.10 LI-COR results for the species in the Paniceae tribe. ................................................................. 87

A.1 RGB images collected for pearl millet, genotype C042, experiment A, between 13 October 2017 and 05 February 2018. ......................................................... 124
A.2 Plant growth curves measured for all species in all angles through the period of 115 days after planting. ................................................................. 125
A.3 RGB images collected for sorghum BTx623, experiment B, between 01 January 2018 and 30 January 2018. ................................................................. 125
List of Tables

2.1 List of species and sequence identifiers used in this study. 41
2.2 Summary of symbols 42
3.1 Published reference genomes for grass species within the PACMAD clade. 47
3.2 Summary statistics for raw and processed long read sequence data generated from each of the three target species. 55
3.3 Alignment rates of consensus transcripts generated from each of the three target species to the sorghum gene space. 55
4.1 List of enzymes with syntenic orthologs investigated in this study. 72
4.2 List of enzymes with fast evolving tree branches and species. 75
A.1 Summary of species information. 120
A.S1 Experimental layout. 126
A.S2 Summary of watering data of experiment B. 130
1: Literature Review

1.1 Overview and importance of C₄ photosynthesis

C₄ photosynthesis evolved about 30 Mya as a response to a drop of CO₂ levels in the atmosphere, which conferred an advantage for those plants to diversify and dominate different environments.¹ About 30% of the terrestrial productivity in the world is obtained from C₄ photosynthesizing plants,²,³ and represent more than 20% of human primary consumption directly or indirectly (via animal products).⁴,⁵ Moreover, C₄ grasses are responsible for about 18% of the world productivity,⁶,⁷ which encompasses crops such as maize and sorghum.

The C₄ photosynthetic pathway was first described over 50 years ago in sugarcane, when Hatch and Slack noticed that oxaloacetate (OAA), a 4 carbon compound, was the result of the first CO₂ fixation step of the pathway.⁸ However, over 20 years prior to the characterization of the C₄ pathway, Rhoades and Carvalho (1944)⁹ noticed that the bundle-sheath cells of maize leaves were responsible for the production of starch, which occurs in the mesophyll cells in C₃ plants. Even though C₃ plants are equally or more efficient than C₄ plants in temperatures around 25°C,³,⁶ other differences between C₄ and C₃ plants confer advantages of the C₄ over the C₃ pathway. C₄ plants can grow in environments that are usually inhabitable for C₃, such as in saline soils and arid environments.¹⁰ Such advantages are only possible due to anatomical and physiological adaptations present in C₄ plants.

The comparison of leaf cross sections of both C₃ and C₄ plants shows some critical anatomical differences. Both plants have leaves that contain mesophyll cells (cells located
in the leaf between both epidermis layers) and bundle-sheath cells (the cells surrounding leaf vasculature). However, in most C\textsubscript{3} plants the mesophyll cells (MC) do not show a specific organization in the leaf or around the bundle-sheath (BS) cells. On the other hand, C\textsubscript{4} plants show a different type of anatomy where the MC are arranged around the BS cells; such cellular organization is called Kranz anatomy.\textsuperscript{11,12} Because of the differences in the anatomy, there are also particular aspects related to their physiology that differs from one type of plant to the other.

In C\textsubscript{3} plants, both carbon fixation and starch production occur in the mesophyll cells. However, in C\textsubscript{4} plants those processes occur in two different cell types. In C\textsubscript{4} photosynthesis, carbon fixation happens in two steps, first in the mesophyll and then in the bundle-sheath cells. In the mesophyll cells, C\textsubscript{4} plants fix carbon in the form of OAA, which is then used to form either malate or aspartate. After that, these two dicarboxylic acids are transported to the bundle-sheath cells, where they are decarboxylated and the released CO\textsubscript{2} is refixed by the Calvin-Benson cycle to generate sugars and starch. This process helps the plant increase the CO\textsubscript{2} ratio in the bundle-sheath. In both photosynthesis types the carbon fixation into the Calvin-Benson cycle is performed by RuBisCO (Ribulose Bisphosphate Carboxylase/Oxygenase).\textsuperscript{13}

Although RuBisCO has an important role in CO\textsubscript{2} fixation, this enzyme can also bind to oxygen. When this happens, the enzyme undergoes a process called photorespiration. This process has a high cost for the plant because it requires energy, releasing CO\textsubscript{2}, lowering the efficiency of the Calvin-Benson cycle\textsuperscript{14} and, in the end of the pathway, no sugars are generated. Because C\textsubscript{3} plants lack the kranz anatomy, both carbon fixation and photosynthesis occur in the mesophyll cells, which increases the chances RuBisCO will fix oxygen.\textsuperscript{12} Higher temperatures make this scenario even worse because it increases the chances RuBisCO will bind O\textsubscript{2} and undergo photorespiration. Due to the fact that both carbon fixation and starch production steps occur in the mesophyll in C\textsubscript{3} plants, they present higher rates of photorespiration than the C\textsubscript{4} counterparts, specially
in hot environments, where RuBisCO is more likely to bind $O_2$.\textsuperscript{15} Also, carbon fixation happens in two steps in $C_4$, hence they are able to concentrate $CO_2$ around RuBisCO, lowering the rate of photorespiration.

In spite of the fact that photorespiration is a process that results in lower Calvin-Benson cycle efficiency and yield loss, this process is still important for both $C_3$ and $C_4$ plants. The knockout of genes related to photorespiration can lead to lethality in both types of plants.\textsuperscript{16} Even in $C_4$ plants, which are able to decrease photorespiration levels due to the $CO_2$ accumulation around RuBisCO, this process is very vital for the plant. In maize, mutant plants lacking a key enzyme for the normal functioning of the photorespiratory chain resulted in the death of seedlings showing that $C_4$ photosynthesis is dependent on this process.\textsuperscript{17} Although it was shown that photorespiration seems to be important to avoid photoinhibition in $C_3$,\textsuperscript{18} in maize it is proposed that this process is crucial for keeping low levels of glycolate in the leaf cells, which can be lethal to the plant in high concentrations.\textsuperscript{17}

Even though $C_4$ photosynthesis has been known for about 50 years, there is still a lot to be learned. Several adaptations are needed for a plant to perform the $C_4$, such as: anatomical changes with presence of the Kranz anatomy, physiological and metabolic adaptations which allow the carbon fixation to occur in two different cell types. Also, the $C_4$ pathway has evolved independently over 60 times in plants,\textsuperscript{19} which increases the challenges for a better understanding of the pathway, especially because $C_4$ photosynthesis can be performed in different ways. In order to increase our knowledge about this pathway it is crucial to develop different methods to unravel the multiple aspects on which $C_4$ plants rely. Therefore, the main goal of this review is to emphasize the different approaches applied to study the complexities of $C_4$ photosynthesis and its specific changes due to parallel evolution.
1.2 Parallel evolution and specific changes

C₄ photosynthesis can be grouped into three broad categories, depending on which decarboxylating enzyme is used to provide CO₂ to be fixated by RuBisCO. This is despite the fact that this photosynthetic pathway has undergone parallel evolution and evolved independently in several groups of flowering plants, including 22-24 times in grasses alone.²⁰ The three different decarboxylating enzymes are: NAD-ME (NAD-malic enzyme), NADP-ME (NADP-malic enzyme) and PEP-CK (PEP carboxykinase).²¹ Such subtypes appear multiple times within grasses without a phylogenetic pattern, as well as in eudicots and monocots in general. These three subtypes also differ in chloroplast localization in the BS cells. In NADP-ME plants the chloroplasts are centrifugally arranged in the BS cells, while in the NAD-ME plants these chloroplasts are centripetally arranged and they are also very close to high numbers of mitochondria. Finally, in PEP-CK plants chloroplasts and mitochondria are evenly distributed.²¹ These differences in chloroplast and mitochondria localizations are consistent with the fact that in the NADP-ME the mitochondria is not part of the pathway, while in both NAD-ME and PEP-CK the mitochondrial pathway is part of photosynthesis.¹³,²¹ Moreover, each one of these C₄ photosynthetic subtypes use slightly different genes and enzymes to perform C₄ photosynthesis (Figure 1.1).

Gene duplications and subsequent neofunctionalizations were a major evolutionary force that lead to the evolution of the C₄ pathway.²³,²⁴ It is known that the enzymes used to perform C₄ photosynthesis are also present in C₃ plants (i.e. PPDK, CA, PEPC), however they have different functions, unrelated to photosynthesis.²⁵ The duplication events and the multi-gene families allow different genes of the same family to evolve under distinct evolutionary rates. So, at least one of the genes in a family can keep its original function while the other copies can neofunctionalize and encode enzymes with different roles. Therefore, the main explanation for the evolution of the C₄ pathway is
that those genes were co-opted from preexisting C_{3} genes. Because of these duplication events, C_{4} metabolic networks are more redundant than C_{3}, which means that C_{4} is more resilient to the removal of enzymatic reactions. This networks resilience probably led to other metabolic differences across the C_{4} subtypes. For instance, there are C_{4} plants that perform more than one subtype pathway, as seen in maize which performs NADP-ME subtype that also uses supplemental PEP-CK activity, and Setaria italica, which can perform both NAD-ME and NADP-ME pathways in different developmental stages.

Since the C_{4} pathway evolved as a consequence of several gene duplications and neofunctionalizations, gene duplication events were characterized within different groups of orthologous genes in two independent origins of C_{4} photosynthesis in grasses. In about half of those events the same gene copy was duplicated and occurred in parallel in different species. This result is consistent with previous studies, suggesting that
genes are not randomly co-opted into C₄ photosynthesis. However, some gene families may recruit different gene members into C₄ photosynthesis, such as DIT2 (a dicarboxylate transporter) in maize, which did not use the same gene copy as the one used when the C₄ pathway first evolved.³⁰

Despite its independent origins, C₄ photosynthesis seems to be recruiting the same gene member from multi-gene families into the pathway. Analysis of enzymes involved in the NADP-ME subtype (such as NADP-ME, PEPC, PPDK and CA) (Figure 1.1A) support this hypothesis.³¹ Also, genes that belong to families of enzymes co-opted into C₄ that have higher expression in C₃ plants were better suited and co-opted for C₄ photosynthesis.³⁰ These results support previous findings that the co-option of a gene from a multi-gene family may not be entirely random and that the same gene copy is being recruited to perform a new function in C₄ photosynthesis in multiple origins.

In spite of which genes have been co-opted into C₄ photosynthesis and the metabolic differences in each pathway, they also differ in terms of how they respond to environmental stresses and yield production, which can also be a consequence of the parallel evolution. Plants that perform NADP-ME photosynthesis have higher growth rates under elevated CO₂ concentrations compared to the NAD-ME subtype.³⁴ On the other hand, the NAD-ME subtype has higher water use efficiency compared to NADP-ME.³⁵ A study conducted in the Namib desert concluded that NADP-ME plants had the highest CO₂ fixation rates and thrived in regions with higher rainfall, while the NAD-ME subtype had the lowest CO₂ fixation rates and dominated arid regions, and PEP-CK plants had intermediate fixation rates and were more present in regions with moderate rainfall.³⁰ In terms of differences in yield, monocot plants that perform NADP-ME and PEP-CK pathways had similar yields and were both higher than NAD-ME.³⁶ Even though these studies are very relevant to improving our knowledge about C₄ photosynthesis, they are also subject to confounding variables due to different plant growth rates, environments these plants are adapted to and where they were
grown for a certain study. Therefore, these variables may lead to inaccurate results.

The C₄ pathway can be studied under different perspectives, including different gene members from a certain family being co-opted into the pathway, metabolic pathways and enzymatic reactions evolving differently in independent origins and how genes are expressed in each one of those scenarios. Moreover, the methods applied to these studies provide interpretations of C₄ photosynthesis that complement each other. Therefore, the development and application of new methodologies and approaches have proven to be remarkably important to understanding the multiple aspects on which the evolution of this pathway relies.

1.3 Approaches for comparative analysis

When C₄ photosynthesis was first described in 1966 and this pathway started being more studied, the research performed to study the pathway and its enzymes mostly used biochemical characterization, enzymatic assays and carbon isotope discrimination. Despite their relevance, these studies can be very expensive, laborious and usually generate information about a few enzymes or genes at a time.

Also, studying C₄ photosynthesis through a classical genetics approach is very challenging, since knocking out genes related to photosynthesis is lethal to both C₃ and C₄ plants. On the other hand, with the current advances in molecular biology and bioinformatics large amounts of biological data are being generated. Therefore, this review is focused on approaches related to bioinformatics and big data, describing in more details the following topics: a) phylogenetics; b) gene by gene analysis; c) whole genome comparison; d) whole transcriptome comparison and gene regulation; and e) C₄ pathway modeling.
1.3.1 Phylogenetics

Using different genes as markers to study evolution of C₄ photosynthesis can also be very informative. In such an approach, different phylogenies are generated for each gene and then compared across all of them to check if they are consistent with one another. Phylogenetic analysis of different chloroplastic gene markers compared with climate data from over 1,000 grass species suggests that the evolution of C₄ photosynthesis seems to be related to drier environments and to changes in the environment from a closed canopy moist forest to open tropical savanna environment. This outcome is consistent with the hypothesis that the C₄ pathway evolved as a consequence of both low CO₂ concentrations and warmer environments.

The nuclear gene phyB from 97 grasses was used to build a phylogeny and dated the origins of C₄ photosynthesis around 30Mya. Using a similar approach, Edwards et al (2011) used three chloroplastic gene markers (rbcL, ndhF, and intron region between trnK/matK) to build a phylogeny of 531 grass species. With the results of this phylogeny, it was possible to infer that C₄ plants evolved from a C₃ ancestor, since most species in the tree are C₃ and that once a branch evolved to a C₄ pathway it hardly reverted back to a C₃ pathway, and that C₄ photosynthesis evolved 22-24 times in the grasses. Phylogenetic studies are very relevant to better understand the evolution of the C₄ pathway. Also, phylogenetic methods can be used by different approaches, as seen in the next sections.

1.3.2 Gene by gene analysis

In gene by gene analysis one or more genes are studied separately in order to understand the changes and how each gene evolves, using phylogenetic approaches, amino acid substitutions and estimating how fast a gene is evolving. This section is divided into two broad categories: single gene studies in one species, and single gene studies in multiple species.
**Single gene, one species**  Some researchers study the C₄ pathway by knocking out genes to assess their relevance to the pathway. To address the relevance of carbonic anhydrase (CA) in maize, an enzyme that converts CO₂ into bicarbonate, which is then fixed to form OAA, the gene encoding this enzyme had its expression decreased to very low levels. So, the maize mutants were grown in environments with high, normal (ambient) and low CO₂ pressure. The results indicated that in both high and ambient levels the CO₂ assimilation was very similar to wild-type plants. However, under low levels the mutant plants were significantly smaller than the wild-type. In a similar study, Osborn et al (2016) performed the same analysis in *Setaria viridis*, which obtained similar results as the ones obtained in maize. These studies suggest that either CA is not a photosynthesis limiting enzyme in the C₄ pathway under ambient CO₂ levels, or that this enzyme is selected under arid environments, since stomatal aperture under such climate is reduced and plants would need to have better CO₂ usage. Although both studies from Studer et al (2014) and Osborn et al (2016) show that CA is not very relevant to C₄ photosynthesis, Zhang et al (2015) found SNPs in the CA gene, among other genes, and correlated CA to nitrogen and carbon metabolism, which is a very important pathway for plant growth and maintenance of the C₄ cycle.

As an attempt to study the function of photorespiration in C₄ photosynthesis, maize plants with low and no activity of the glycolate oxidase gene GOI were created in order to compare the phenotypes of wild-type plants with heterozygotes and homozygotes for the goI mutation. GOI mutants are unable to convert glycolate into glyoxylate, which is then used for the normal photorespiratory pathway. Although it was not possible to see phenotypic differences across the plants when grown under high CO₂ concentrations, under ambient CO₂ levels the heterozygotes showed a smaller CO₂ assimilation, while the homozygote mutant was dead after 2 weeks. As the main result, it was proposed high levels of glycolate are lethal to the plant, hence photorespiration is an important process to balance the levels of glycolate in the leaf cells.
Although important to characterize enzymatic functions, studies performed in a single gene and species can be very laborious. Also, these studies only provides information about one species at a time, which is unrealistic and slow to be done for every single gene in any pathway, creating a barrier for studying C₄ photosynthesis in a high-throughput manner. One possible solution to make single gene studies faster would be by using multiple species at a time.

**Single gene, multiple species** Despite its limitation to one gene, using multiple species to understand the evolution of a single gene can provide insights about how that gene has changed and which mutations are shared across species. For instance, this approach can be very valuable for checking for parallel mutations in different species and identifying similar and/or different mutations across C₄ plants.

Christin et al (2007) sampled 111 grass species, amplified and sequenced PEPC genes from genomic DNA and performed phylogenetic analyses with the amino acid sequences obtained. Building the phylogeny of PEPC genes using all three positions of the gene showed all C₄ PEPC genes have a common ancestor. However, analyzing only the third position of codons revealed that C₄ PEPC genes were polyphyletic. These results indicated that the PEPC gene evolved 8 times independently, and identified key amino acid position that had specific changes in C₄ plants that were absent in C₃ plants.

In a similar study, PEP-CK genes from 57 grass species were studied using phylogenetic analyses and a positive selection test. Like the PEPC genes, the C₄ PEP-CK genes also showed amino acid sites under significant positive selection with amino acids present in those positions in C₄ species and absent in C₃ counterparts. Moreover, they noticed that the branches leading to C₄ PEP-CK copies were evolving faster than the other gene copies.

In the genus Flaveria, the genes of the large and small subunits of RuBisCO (rbcL and rbcS, respectively) were analyzed in C₃ and C₄ species and tested for positive selection.
Also, the amino acid changes in both genes were compared with the biochemistry of the enzyme to support the findings for the sites under positive selection. The main conclusion was that both subunits were under positive selection in the C₄ relatives.⁴²

Overall, the studies mentioned in this section were relevant to better understanding the evolution of C₄ genes separately. However, like the use of single gene in one species approach, studies involving multiple species also represent a barrier for understanding the evolution of multiple genes and become a limiting step for further unraveling the evolution of the C₄ pathway as a whole.

Finally, gene by gene analysis can provide several insights in terms of how different genes evolve, how important they are to the C₄ pathway and/or how and when the pathway appeared. However, the number of genes studied can be limited to a small number. Therefore, using whole genome data can help increase this number and address problems concerning other genes that might also be involved in C₄ photosynthesis.

1.3.3 Whole genome comparison

Whole genome comparisons can help identify new genes and homology relationships, provide a broader view of multi-gene families and their evolution, and check if the genes being compared across genomes share the same genomic location (syntenic genes) or not.²⁴,²⁸,⁴³ Some studies focus on identifying genetic targets to explain an evolutionary or ecological phenotype, such as the different C₄ subtypes and their environmental adaptation. During an Ecological Genomics symposium in 2007, Matthew Rockman proposed the term "reverse ecology" to describe the approaches used to identify the genetic traits under selection to explain the evolution of a phenotypic or ecological trait.⁴⁴

The use of reverse ecology, along with syntenic genes for identifying C₄ genes, can result in the identification of new genes previously not described as involved in the pathway, searching for genes under adaptive selection. In 2016, Huang et al (2016)²⁸ compared the genomes of 6 grass species, where 3 of them were C₃ plants and performed
A synteny analysis, testing the syntenic genes under positive selection by calculating the rates of non-synonymous/synonymous substitutions and testing different phylogenetic scenarios removing species from the analysis. The null hypothesis of the study was that \( C_3 \) and \( C_4 \) branches evolved at the same rate, while the alternative hypothesis stated that \( C_4 \) had an independent rate from the \( C_3 \) branches. This study yielded in a list of 88 genes where \( C_4 \) branches were under positive selection, considering them as candidate genes involved in the \( C_4 \) pathway. Such a study is very useful because it provides a start to further enzymatic and functional characterization studies.

Although synteny analysis is widely used to study gene homology relationships, it is not the only method to identify clusters of orthologous genes. OrthoMCL is a software that allows the comparison across multiple genomes, clustering orthologous genes based on BLAST searches and clustering algorithms.\(^{45}\) Aubry et al (2014)\(^ {46}\) have used this method to improve their ability to identify clusters of homologous genes and study their expression comparing maize and Cleome gynandra, a monocot and a dicot, respectively. In this study, however, the highest accuracy of abundance estimates was around 80%, which was still considered low. Also, this method can be complex because of the multiple parameters and programs involved in the software. Further, such method does not allow the inference of gene localization, which can add more noise to the analysis, since genes that do not share the same location are less likely to have conserved function than syntenic ones.\(^ {43}\)

A comparative genomics analysis across maize, sorghum (both \( C_4 \)) and rice (\( C_3 \)) was performed by checking gene synteny, and phylogenetic analysis to identify the origins of gene copies to infer adaptive evolution. With these analyses it was possible to identify several \( C_4 \) genes that underwent positive selection, gene copies that neofunctionalized and also the genes that were found in the same genomic location in these species.\(^ {24}\) Syntenic genes are important for studying \( C_4 \) photosynthesis because such genes usually have conserved functions.\(^ {43}\) So if syntenic genes do not perform the same function
anymore it is likely that this change was not just random.

Most studies focus on performing genomic analysis in multiple species. However using different varieties within a single species can also provide valuable information about genes involved in the C₄ cycle. Another way to investigate whole genome changes related to C₄ photosynthesis is by performing a genome-wide association study (GWAS). In this analysis phenotype data is collected and associated to single nucleotide polymorphisms (SNPs) throughout the genome. Applying a GWAS approach, Zhang et al. (2015) genotyped 5,000 maize inbred lines and their measured metabolites to find SNPs linked to known C₄ genes related to carbon and nitrogen metabolism. This study resulted in identifying SNPs close to genes with unknown function that were related to nitrogen and carbon metabolism. Although the results help identify specific genes related to a certain metabolic trait, it can also be limited by the fact that at least hundreds of plants are needed to provide the statistical power to make any assumptions about genes linked to a certain trait.

In summary, whole genome analysis can identify genes with similar functions, genomic locations and different evolutionary rates. However, it is still possible that similar genes do not share similar expression patterns, which implicates in inferring function to a gene that has very low expression or is not expressed at all. Therefore, further transcriptome analysis can help bypass such limitation, providing new insights into the evolution and expression of different genes.

1.3.4 Whole transcriptome comparison and gene regulation

Transcriptome and gene expression data can provide insights about genes that show a similar pattern and how highly/lowly expressed that gene is compared to its expression in another species. As one of the possible consequences of the parallel evolution, C₄ regulatory networks are also subject to some changes in gene expression and regulation. In order to perform such kind of analysis it is important to generate expression data
from multiple species. With the advance of sequencing techniques and the lowering costs to sequence RNA, several groups are now able to generate a large number of data, usually available online for free. Also, in order to have a better designed transcriptomic comparison, it is ideal that the data for the species are collected and generated in similar conditions, hence avoiding comparing gene expression under different conditions or developmental stages.

Recently, a method for comparing gene expression in rice and maize by analyzing comparable leaf developmental stages has shown that both species share similar expression patterns of photosynthesis related genes in the same leaf segments. Also, the method allowed the identification of expression of photorespiration genes, which is high in mature rice plants, but are only expressed in immature maize leaf segments. C₄ plants share similar regulatory factors in syntenic genes, but they are missing in rice genes, showing that evolution of the C₄ pathway required novel changes in gene regulation. Despite the differences among regulatory factors, a comparison between maize (C₄) and *C. gynandra* (C₃) revealed that, even though the common ancestor of these two species has diverged about 140 Mya, they both share transcription factors related to C₄ photosynthesis and cell specificity. This last study shows that despite the evolutionary divergence of two species, they still share common regulatory factors, likely to be essential for both types of photosynthesis.

In 2016, Covshoff et al. generated transcriptomic data to study the adaptive mechanisms of *Echinochloa glabrescens*, a C₄ plant, compared to paddy rice, since *E. glabrescens* is a weed that compromise paddy rice growth. Not only this weed is adapted to paddy conditions, but also shows differences in the C₄ pathway. Although transcript abundance of both NADP-ME and PEP-CK enzymes are higher in *E. glabrescens*, CA and PEPC have higher abundance in rice. Despite some changes in the expression of C₄ genes, the transcription factors with high expression show similar patterns in maize and *S. italica* as well.
Another relevant aspect of transcriptomic analysis is to address questions about gene specificity and where it is expressed. The comparison of expression data from maize and *Setaria viridis*, which represent two independent origins of C₄ photosynthesis, shows that both species have similar gene expression patterns in both mesophyll and bundle-sheath cells. However, those two species represent the same C₄ subtype, NADP-ME with some level of PEP-CK activity. Another cell specificity expression study was performed in eudicots, *Gynandropsis gynandra* (C₄) and *Tareneya hassleriana* (C₃), which concluded that expression of C₄ genes was higher in the guard cells of *G. gynandra* than *T. hassleriana*. This indicated that despite the fact that guard cells are related to controlling stomatal aperture, these cells also have higher expression of C₄ photosynthesis genes compared to their C₃ counterparts, also indicating the presence of two different regulatory networks in stomata of C₄ leaves. Such difference highlights the fact that not only the C₄ pathway itself is different than C₃, but also suggests that there are also significant differences in the way the stomata of both photosynthesizing pathways can be related with the fact that C₄ plants have higher CO₂ fixation rates.

Plant groups with closely related C₃ and C₄ relatives can improve our knowledge about more recent evolutionary events. The genus *Flaveria* is a plant genus that comprises both C₃ and C₄ species. Throughout leaf maturation it was detected that there is an increase in expression of C₄ genes in the C₄ species, a pattern not seen in the C₃ counterparts. However, genes related to photorespiration have increased expression in C₃ species compared with the C₄ ones. This result is consistent with the fact that C₄ plants have lower rates of photorespiration. Leaf anatomy is another aspect of C₄ plants that is subject to different gene regulatory networks. In maize the SCARECROW transcription factor (TF), along with other TFs, was indicated as one of the genes responsible for regulating the development of kranz anatomy in leaves. However, in C₃ plants this gene is responsible for the normal development of the shoot.
1.3.5 **C₄ pathway modeling**

Despite all the advances in sequencing methods and big data, such as genomics and transcriptomics data, *in silico* modeling analyses are also being developed and used to address biological questions. In C₄ photosynthesis research computational modeling of the pathway has proven to be a valuable tool to predict different outcomes of the metabolic processes, simulations of stress conditions and mixing different subtype pathways. With the improvement of computational power, metabolic modeling of the C₄ photosynthetic pathway is increasingly being used to overcome some limitations of molecular, metabolic and enzymatic characterization of the pathway. With such approach, it is possible to predict different outcomes of metabolic reactions based on equations that model different aspects, such as: substrate and product concentration; and enzymatic kinetics and abundance, for instance.

One of the approaches used for modeling is the construction of metabolic networks, where the network is composed by nodes (say A and B), which represent different metabolic reactions, and the edges represent a connection between the two reactions in case the final product of reaction A is a substrate for reaction B. This way it is possible to see communities of reactions that share the same product/substrate. This analysis has been applied along with prediction and modeling of different gas concentrations, enzymes, and knocking out genes related to the photosynthetic pathway to infer possible outcomes of C₃ and C₄ pathways. The simulation of different environments and knockouts showed that C₄ networks are more resilient to mutations and/or environmental changes compared to the C₃ networks. This outcome is consistent with the fact that many C₄ genes have undergone duplication events, which could create gene redundancy and help maintain the normal metabolic functions.

Another possible application of metabolic modeling is to discover new metabolic reactions using gene expression data to infer the reaction rates of the pathway. For instance, previously published maize leaf metabolic models were used to unravel the
reactions that happened in the mesophyll and bundle-sheath cells and making a correlations with gene expression. The main conclusion of this study was that metabolic reactions correlated with each other may shape the co-expression patterns of the genes involved.53

Besides describing metabolic reactions specific to different cell types, correlations between gene expression and metabolic pathways can be used to address gene expression in different developmental stages of leaf tissue. In a study performed in 2016,29 different sets of data were generated for S. italica to make a metabolic reconstruction, such as RNA, protein and metabolite extractions and analyses. By merging all the data a metabolic model was retrieved by identifying S. italica transcripts in their previously published C_4 model, called C4GEM,54 doing a BLAST search of maize genes that map to the transcripts found. Finally, the metabolic data was merged with the expression data from young and mature leaf tissue. The conclusions drawn from the results was that the expression levels of genes from different metabolic pathways are very different and that S. italica, classified as NADP-ME subtype, can perform both NAD-ME and NADP-ME pathways in different developmental stages.

The correlation of leaf gene expression and enzyme activity levels from Megathyrsus maximus, sorghum (both C_4), Brachypodium dystachyon and rice (both C_3) was also studied.55 This comparison yielded in the characterization that M. maximus, classified as PEP-CK subtype, actually has shown some level of NAD-ME expression as well. Some studies question the current most accepted idea that there are three different biochemical subtypes of C_4 photosynthesis.47 Using a similar methodology as a previous study55 to build a metabolic model for the C_4 pathway, Wang et al (2014)47 also created models with different possibilities of mixed pathways, addressing their viability and yield. As conclusions they indicated that the PEP-CK pathway is not viable to exist without another supporting pathway, supporting the idea that there are only two main subtypes of C_4 photosynthesis that either use NAD-ME, NADP-ME or a mixture of either
pathway with a supplementary PEP-CK pathway, as seen in maize and S. viridis. However, a recent study suggests that Urochloa fusca can perform a purely PEP-CK cycle without use of any extra decarboxylating enzymes.

1.4 Engineering the C₄ pathway

One of the main goals of crop scientists is to be able to engineer a C₄ pathway into a C₃ crop, such as rice, tobacco and wheat. However, engineering the C₄ network is very challenging because different genes are regulated in specific ways and have different cis- and trans-factors that regulate their expression in specific tissues and pathways. Further, previous studies have identified transcription factors (TFs) common to different C₄ species, although there are TFs that are specific to controlling C₄ photosynthesis in each species.

Overall, among the difficulties to engineer a C₄ pathway are: transferring the C₄ genetic networks into a C₃ plant, transforming the leaf anatomy of a C₃ leaf into developing kranz anatomy and recreating the C₄ metabolic networks in a C₃ plant. Despite the challenges of engineering a C₄ pathway, over time several genes and enzymes from the pathway are described including where they are expressed, which improves knowledge about C₄ and provides insights about genes that should be tested, or engineered into C₃ crops.

Different genes have been tested to create mutants of rice plants with C₄ related genes through genetic engineering. One of the biggest challenges in performing such experiment is that C₃ plants lack the kranz anatomy in their leaves. A set of 60 candidate genes related to kranz anatomy in maize was tested in a attempt to develop transgenic rice plants capable of developing such anatomy. Different mutants showed abnormal shoot and root anatomies, besides showing major changes in the leaf anatomy. This study was not successful in terms of creating a rice plant with kranz anatomy, which reinforces the difficulty of engineering a C₃ plant to perform C₄ photosynthesis. Even
though a molecular mechanism to induce the formation of the kranz anatomy in $C_3$ plants is yet to be found, a recent study showed that a single developmental step, increasing the leaf vein density, is enough to create the kranz anatomy in plants.\footnote{61} This study of change in leaf density can be supported by the hypothesis that auxin, along with other transcription factors such as brassinosteroids, SHORTROOT/SCARECROW and INDETERMINATE DOMAIN, play an important role in leaf density.\footnote{62, 63}

A good plant candidate to improve our knowledge about $C_4$ photosynthesis is the eudicot $Bienertia$ $sinuspersici$, which is capable of performing photosynthesis within a single cell.\footnote{58} That feature is closer to the way $C_3$ plants perform photosynthesis, since both carbon fixation and the Calvin-Benson cycle happen in the mesophyll cells.\footnote{13} Instead of photosynthesis happening in two cell types, as in maize, sorghum and other $C_4$ plants, in $B. sinuspersici$ the photosynthesizing cells have compartmentalized chloroplasts to perform similar functions as the mesophyll and bundle-sheath cells. In these cells, there are chloroplasts arranged in the periphery and in the center of the cell. Periphery chloroplasts are mainly responsible for the first step of carbon fixation, like the mesophyll, and the central ones are mainly responsible for the pathways related to starch generation, like the bundle-sheath.\footnote{58} So, unraveling how to transform a $C_3$ plant to perform $C_4$ photosynthesis like $B. sinuspersici$ could possibly be less complicated than trying to engineer the kranz anatomy into $C_3$ plants.

\section*{1.5 Concluding remarks}

Even though many researchers are making efforts to study $C_4$ photosynthesis applying new technologies, it is still very challenging to overcome the complexities of how this type of photosynthesis evolved. The study of different genes has provided knowledge about the evolution of multi-gene families, identification of specific mutations related to $C_4$, and even dated when this pathway evolved.\footnote{1, 20, 31, 32} Although a lot of progress has been made in the field, questions about whether or not more $C_4$ related genes exist, besides
the known ones, and what regulatory elements control the different aspects of the C₄ pathway remain unresolved.

Also, the development of technologies involving generation of big data has helped in identifying new genes that were not previously described as part of the C₄ pathway.²⁸ Further, modeling studies indicate important features and differences of the subtypes of photosynthesis, which has lead to a better understanding of how this pathway can be improved.²⁶,²⁹,⁵⁹ However, the approaches presented here have different weaknesses and strengths, which reinforces the need to develop new ways to study C₄ photosynthesis.

In terms of improving the pathway, engineering a better C₄ pathway is a promising alternative to increase crop yield, especially because the current food production curve it will not be increased enough to feed the estimated population by 2050.⁶⁴ In summary, increasing the efforts and approaches to study C₄ photosynthesis with the use of modern methodologies has proven to be very important for improving our knowledge of the pathway from several different perspectives.
2. Integrating phylogenetic and network approaches to study gene family evolution: the case of the *AGAMOUS* family of floral genes

2.1 Introduction

Advances in sequencing technology have lead to dramatic expansions in the number of sequenced genes within most gene families, both through the use of whole genome or whole transcriptome sequencing, or through broader taxon sampling. Gene families are generally studied through the use of phylogenetic approaches in order to identify closely and distantly related sequences, as well as to classify divergence between gene copies into those resulting from speciation (orthology) or gene duplication (paralogy). Thus, phylogenetic approaches are widely employed to study how sequence divergence can lead to divergence of structure and/or function. When coupled with genome-context information, this approach can provide insightful understanding of gene regulation and function.

For instance, it is well-known that orthologous genes conserved at syntenic locations in the genome are more likely to exhibit conserved regulation and function than genes at nonsyntenic locations. On the other hand, the prevalence of whole genome duplications in plants poses challenges to the study of gene family evolution using exclusively phylogeny-based methods due to the diverse outcomes of duplicated genes. Whole genome duplications produce syntenic paralogs that can be reciprocally lost, sub or neofunctionalized, or even retained in the same functional roles as a result of relative or absolute dosage constraints.

A fundamental assumption of any phylogenetic reconstruction is that the observed
changes occur exclusively through a hierarchical bifurcated branching process. This model is certainly a good representation of a major evolutionary force (i.e., descent with modification); however many will argue that it fails to capture the diversity of evolutionary processes which shape the gene content of extant species.

One way to address the complexity of evolutionary processes is to apply network approaches to address questions related to cell organization and functioning, human diseases relationships and plant gene function prediction. Network approaches have also been successfully applied to study fungi evolution based on enzymes related to the chitin synthase pathway. Recently, Carvalho et al. have used a network-based approach to address the origin of the mitochondria, providing a new perspective on the study of mitochondrial evolution.

Network-based approaches can overcome some of the limitations of phylogenetic methods. For instance, these approaches do not require the assumption of a hierarchical bifurcating framework and therefore may be capable of dealing with more complex biological patterns and phenomena. Networks are generally less precise in their ability to reconstruct the divergence points of different groups within a gene family, however, they may be able to capture additional insight into function evolution and divergence using information which might be lost in phylogenetic reconstructions.

In this study we compare the information gained from conventional phylogenetic analysis and a network-based approach using a well characterized subfamily of floral transcription factors, the AGAMOUS floral genes. The AGAMOUS gene subfamily comprises MADS-box transcription factors and is involved in important aspects of flower and fruit development. Among angiosperms (flowering plants), the AGAMOUS subfamily is traditionally divided into the C and D lineages. C lineage genes include the closest relatives of the Arabidopsis thaliana AGAMOUS (AG) gene in all angiosperm, as well as close relatives of SHATTERPROOF (SHP) gene, present exclusively in core eudicots.
On the other hand, the D lineage includes angiosperm *SEEDSTICK (STK)* genes. The C/D split likely occurred after the split between gymnosperms and angiosperms. Thus, gymnosperms usually carry a single gene crop from the AGAMOUS subfamily. While D lineage genes are usually related to ovule development, C lineage genes have been implicated in stamen and carpel development. Particularly in core eudicots, *SHP* genes have also been shown to be involved in fruit development and ripening.

This gene subfamily has been extensively studied and mutant characterization has provided insights into their functional roles in carpel, ovule and fruit development as well as floral meristem termination. The AGAMOUS subfamily has undergone several instances of duplication followed by neo and subfunctionalization throughout its evolutionary history in angiosperms (as reviewed) and understanding the evolutionary history of this group has proven challenging as a result of low support for deep nodes on the tree.

Here we propose using a similarity-based phylogenetic network approach. The phylogenetic network methods used here does not require the assumption of a scale-free topology, or the need to calculate gene correlation based on expression data, which makes the approach used more straightforward. Also, the approach used here does not rely on an existing tree in order to generate the networks, as with most phylogenetic networks. Overall, both the phylogeny and network results showed consistent clustering of the gene families. However our results suggest that the network approach was less affected by sequence divergence. We demonstrate that a combination of both methods may provide additional insight into evolutionary events and functional divergence within gene families.
2.2 Methods

Sequence search and multiple sequence alignment

C and D lineage AGAMOUS nucleotide sequences were retrieved on Phytozome (https://phytozome.jgi.doe.gov/pz/portal.html) and NCBI. Species of origin and accession numbers for each sequence included in this analysis are provided in Table 2.1. A multiple sequence alignment was performed using the ClustalW\textsuperscript{99} alignment tool within Geneious\textsuperscript{®}v7.0.4,\textsuperscript{100} based on translated nucleotides. Further refinements were made manually, using translated sequences as a way to guide manual curation. Manual curation of the multiple sequence alignment was performed using a codon preserving approach and taking into account domains and motifs previously described in the literature.\textsuperscript{88} Unalignable regions were removed prior to further analysis. The final multiple sequence alignment included 549 nucleotides. The alignment statistics obtained from HMMSTAT, from HMMER3 package,\textsuperscript{101} were: eff\_nseq = 2.72, M = 531, relent = 0.45, info = 0.45, p relE = 0.31, compKL = 0.02. jModelTest 2.1.1\textsuperscript{102} was used to estimate the best-fit evolutionary model of nucleotide evolution. A protein multiple sequence alignment was also performed with the same sequences and used in downstream phylogenetic analysis.

Phylogenetic analysis

Maximum likelihood analysis was performed using PhyML 3.0 (http://www.atgc-montpellier.fr/phyml/)\textsuperscript{103,104} with the TN93 model\textsuperscript{105} a gamma distribution parameter of 1.107. Bootstrap support was calculated based on 100 iterations. The most likely tree was computed based on the PhyML estimated parameters: transition/transversion ratio for purines of 2.541, transition/transversion ratio for pyrimidines of 4.342, and nucleotides frequencies of f(A)= 0.33406, f(C)= 0.20359, f(G)= 0.24537, f(T)= 0.21698. A ML tree of the protein sequence multiple sequence alignment
was also performed on PhyML 3.0 using the LG model of amino acid substitution.

**Obtaining identity matrix**

A pairwise distance matrix, based on a nucleotide multiple sequence alignment of the 93 sequences was calculated using MEGA7. Even though the length of the final alignment obtained was 543 positions, removal of gaps and missing data was performed in order to calculate the distance matrix, resulting in a final set of 372 informative positions in the final filtered dataset. The number of base substitutions per site between sequences was calculated using the Maximum Composite Likelihood model. To obtain the identity value of the sequence pairs, we subtracted 1 from the distance value of every term of the distance matrix to finally obtain the identity matrix.

**Network analysis**

Once the gene identity matrix was generated, a set of 101 networks were created based on the identity threshold between sequence pairs (1 network for each threshold, 0% through 100%), which is represented by the parameter $\sigma$. In each network, each nucleotide sequence is represented by a single node. Two nodes (say $i$ and $j$) are considered connected if the identity threshold is greater than a $\sigma$. The networks were represented in the format of an adjacency matrix $M(\sigma)$, where the matrix elements $M_{ij}$ (pairs of sequences) were either 1, if they were connected, or 0, if they were not connected. Then, neighborhood matrices $\hat{M}(\sigma)$ were built for each one of the $M(\sigma)$. Each element $\hat{m}_{ij}$ from $\hat{M}(\sigma)$ represents the number of steps in the shortest path connecting two nodes $i$ and $j$. Whenever two nodes are not connected and belong in different clusters, $\hat{m}_{ij} = 0$. A neighborhood matrix shows the number of edges connecting two nodes in the network. The neighborhood matrices were later used to calculate the network distance $\delta(\sigma, \sigma + \Delta \sigma)$ between the pairs of successive networks (in this case $\Delta \sigma = 1$), in order to find the network with the most meaningful biological information, as
previously described. Further description of the symbols used here is in Table 2.2.

GePhi was used to visualize and further interrogate the networks. The modularity calculation from GePhi, based on and resolution from was used to classify individual nodes into communities.

To summarize the network approach applied here, we describe the main steps performed:

1) Alignment of gene sequences;
2) Calculation of genetic distances and generation of identity matrix;
3) Calculation of network distances;
4) Identification of best $\sigma$;
5) Network generation and analysis under most informative $\sigma$ value.

The proposed approach used here requires less than 10 seconds to run on an Acer Intel(R) Core(TM) i7-6700 CPU @ 3.40GHz for all datasets tested to date (< 100 sequences). The scripts used here can be found on GitHub (https://github.com/deCarvalho90/network_analysis) and the software with a graphical interface is available in Goes-Neto et al (2018).

2.3 Results

Phylogenetic Analysis

The Maximum Likelihood phylogeny of AGAMOUS genes presented in Figure 2.1 is consistent with the topology previously published studies of the AGAMOUS gene family. The most likely nucleotide tree had a log likelihood score of -20654.546. The ML protein tree had no support for main clades, and therefore was not used in subsequent analysis (data not shown).

Gymnosperm AGAMOUS genes (here termed C/D homologs) form a paraphyly at the base of the unrooted tree. An initial duplication event separates C and D lineage angiosperm genes, and likely occurred in the common ancestor of angiosperms. Basal
angiosperm C lineage homologs, although clustering with D lineage genes, exhibit expression patterns, and likely function, similar to that of core eudicot C lineage genes. D lineage genes form a monophyletic clade that includes all other angiosperm species included in this study.

Monocot D lineage genes appear as a paraphyly at the base of the D lineage clade, however the relationships among D lineage genes otherwise are largely consistent with known species relationships. The relationships of C lineage genes are more convoluted. The base of this subtree is a polyphyly including monocot, basal eudicot and core eudicot genes. At the base of the core eudicots, a second duplication event resulted in the split of the AGAMOUS and PLENA/SHATTERPROOF (SHP) lineages. A third duplication, likely at the base of the Brassicales, resulted in two copies of SHP genes in this group (SHP 1 and SHP2) (Figure 2.1).

Basal angiosperm C lineage genes form a group that diverges after the gymnosperm C/D lineage, but before the angiosperm C/D lineage split. The artificial polyphyletic group of the paleoAGAMOUS includes monocot and basal eudicot sequences. While the basal eudicot group with other core eudicot AGAMOUS genes, monocot paleoAGAMOUS genes share a most recent common ancestor with D lineage genes. It is important to notice, however, that the low branch support in many areas of the AGAMOUS gene tree poses challenges to the interpretation of the evolutionary relationship between clades.

Network Analysis

The network distance graph showed its highest peak at 75% identity, which means that the network generated at that peak is the most distant from the others (Figure 2.2A). Also, it means that the network presents a clear community structure with relevant evolutionary information. Despite the fact that the network with the biggest distance was obtained at 75% identity, the community structure was already too fragmented to answer questions about the evolution of the gene families analyzed in the phylogeny.
Figure 2.1: Phylogenetic tree of the AGAMOUS family genes. Main functional groups are highlighted in black boxes along the tree.

(Figure 2.1A). Even though the network obtained at 75% was too fragmented, the network still provided relevant information about the functional divergence of the genes. However, we wanted to see how the community structure would behave in a scenario closer to the phylogeny. In order to do so, we had to find the network where all
sequences were connected in a way that it would still be possible to retrieve a community structure. A similar situation occurred in Carvalho et al (2015), and the problem was solved by analyzing other networks in different peaks. Here we attempted to solve this problem by analyzing the network at 51% in order to find the last network where all sequences were connected. However, it was not possible to see a clear community structure in this network due to the high degree of connectivity between nodes (Figure 2.S1B). Finally, in this study we focused mainly on the network obtained at the identity threshold 67, which meant that two sequences had to have an identity value of 67% or higher to be connected. The choice of the network threshold was based on the fact that all sequences in this study were connected, with exception of the outgroup sequences, which reflected a scenario similar to the phylogeny.

After applying the modularity calculation (see methods) in the 67% network, it was possible to see the emergence of the community structure of the network, containing five communities (C1-C5) (Figure 2.2B). Each one of the communities mainly cluster genes that have similar functions. In C1 3 out of the 5 nodes from Gymnosperm C/D homologs are connected. Even though the 5 nodes are not connected, this result was expected due to the fact that they are part of the most distant outgroup sequences as seen in Figure 2.1. In C2, on the other hand, the functions of the nodes are related to AG, paleoAG and basal angiosperm C homologs. This might suggest that the basal angiosperm C homologs have retained a function very similar to the AGAMOUS genes. In C3 the SHP genes are clustered together, but in a different community of the AG genes, also suggesting functional divergence. The genes clustered in C4 comprise the STK genes. Even though the communities were mostly composed by genes with similar functions, three genes exhibited unexpected placements. For instance, the SHP gene from Vitis vinifera (ViviSHP) clustered with other AG genes in C2, instead of with other SHP genes in C3. Similarly, Sorghum bicolor SbAG2, a STK gene, clustered in C5, instead of the expected C4, while Sorghum bicolor SbAG3, a paleoAG gene, clustered in C4, instead of the expected C5.
Lastly, the genes clustered in C5 belong to the monocots paleoAG. This result might suggest that monocot paleoAG genes are evolving under different evolutionary forces than the paleoAG and AG lineages. Finally, we can notice that the grouping obtained by both methods were consistent with one another by comparing Figure 2.2D and Figure 2.1. Also, the results obtained at the 67% threshold are largely congruent with the one obtained for the protein network generated (Figure 2S2), obtained at the 60% threshold (highest peak). However, the protein network showed lower resolution, since it clustered together AG, eudicot paleoAG and SHP genes, while we see a clear separation of SHP from the other genes.

Figure 2.2: A) Network distance graph based on the $\delta(\sigma,\sigma+\Delta\sigma)$ distance. The values for the analyzed networks obtained at 51, 67 and 75% are marked. B) Network obtained at 67% identity. Nodes are colored based on the community they belong to (C1-C5), as result of the modularity algorithm (see methods). The sequences that do not belong to any community are represented as gray nodes. C) Network obtained at 67% identity, colored based on gene function. D) Network obtained at 67% identity colored based on species phylogenetic placement.

Even though the 75% network showed a fragmented community structure for this
study, we can notice that it shows that the STK sequences from maize, sorghum, rice and brachypodium are in a separate community. This information might suggest that STK genes from grasses might be undergoing a functional divergence compared to the remaining STK genes, however, limitations in gene functional annotation does not enable us to further support this inference.

Both the phylogenetic and network based analyses returned largely consistent sets of gene clusters. However the grouping of monocots paleoAG sequences in a separate cluster (C5) than other C homologs from basal angiosperm, basal eudicot and eudicot sequences (jointly clustered in C2) in the network based analysis suggest two testable hypothesis: (i) monocot sequences are undergoing different and independent evolutionary processes when compared to other non-monocot AG homologs, and (ii) non-monocot AG sequences are clustered with euAG genes due to conservation of function.

2.4 Discussion

The use of phylogenetic methods to study gene family evolution has provided vast increases in the understanding of molecular evolution, and the utility of these methods for reconstructing ancestral relationships remains unparalleled. However, in many cases complex evolutionary processes including neofunctionalization, repeated co-option into new biological roles, as has occurred in independent origins of C4 photosynthesis, high birth/death gene families, and reciprocal gene loss following gene or genome duplication may indicate reconstructing phylogenetic relationships may not be the most effective method for identifying genes with equivalent functional roles. Among the contributions of a network approach to gene family studies is the interpretation of the relationships among gene sequences that are not limited to a bifurcating pattern, which is often the case in a phylogenetic framework. A network approach allows for the emergence of patterns that are not seen otherwise. Here we propose the use of network-based
approach which has complementary sets of strengths and weaknesses to conventional phylogenetic methods and tested the contributions of these methods using data from the well characterized AGAMOUS family of floral transcription factors.

For instance, in the phylogenetic tree the non-monocot euAG and paleoAG genes are not clustered with the Basal C homologs. Rather, in the networks we notice that these genes are clustered together suggesting a higher functional conservation between them, which is not seen in the tree. Also, from the tree alone we cannot infer if either euAG or SHP genes neofunctionalized. However, because all the euAG and non-monocot paleoAG are all clustered together with the ancestral C homologs, and apart from the SHP genes, we can infer that the SHP genes neofunctionalized while the euAG and non-monocot paleoAG retained an ancestral function. We believe a combined approach might help with discerning functional and structural evolution in a way that neither methods can provide on its own.

In agreement with the literature, the network based analysis recovered clusters of paleoAG and AG genes from basal angiosperms, basal eudicots and core eudicots, potentially indicating conserved functional roles for the genes included in these clusters despite sequence divergence. In contrast, the position of the basal angiosperms C lineage in the phylogenetic tree lead to uncertain interpretations of conserved or divergent function with respect to the D lineage. The network based approach also separated the STK and paleoAG genes within the monocot lineage, despite the close phylogenetic relatedness of these two gene clades, consistent with reports of distinct functional roles for these two sets of genes in monocots. For instance, paleoAG gene from maize have undergone a duplication event in the common ancestor of maize, wheat and rice which led to subfunctionalization of these genes. They perform functions still related to, but different from, Arabidopsis AG. A similar process also occurred in rice. These differences may the the reason the monocot paleoAG clustered together in the network, but in a different community than the remaining AG gene sequences.
Moreover, genetic networks of the inflorescence meristems can vary a lot between grasses and eudicots, since several changes in these regulatory networks are either only present in grasses, or perform a different function in eudicots.\textsuperscript{120}

However, network-based approaches to studying gene families bring with them their own set of limitations. Some of these are inherent to the particular methodology used here, while others are a result of the relative immaturity of statistical and software tools for applying these methods to the analysis of gene family evolution. For example, a range of statistical methods are widely available for estimating the level of support for individual branches/clades within a given phylogeny, such as jackknife, bootstrap and posterior probabilities.\textsuperscript{121,122} In contrast, methods for calculation of cluster support in a biological context are far less mature, at least for the implementation employed here. The use of sequence identity as a measure of distance, while computationally tractable, also means discarding a great deal of information on the frequency of different types of substitutions at both the nucleotide and amino acid level which can be incorporated into many modern phylogenetic algorithms.\textsuperscript{123}

Figure 2.3 summarizes the contributions and relative strengths and weaknesses of phylogenetic and network based approaches to the study of gene family evolution. We propose that the combination of both methods can provide a better assessment of both functional and historical relationships between sequences than either approach alone.

2.5 Conclusions

Investigating the contributions of a particular network-based approach to the study of the evolution of a well-known family of transcription factor genes involved in floral development supports the idea that network-based approaches, when used in conjunction with phylogenetic methods, can be used to improve our understanding of functional conservation or divergence within gene family evolution. The network-based analysis of gene families used here currently lacks the robust ecosystem of computational
Phylogenetic methods

- Hypotheses of ancestral relationships
- Easy identification of duplication events
- Possibility of dating of evolutionary events

Network approaches

- More robust account of gene family evolution:
  - Hypotheses of functional diversification
  - Hypotheses of distinct evolutionary process in action

Independent data to support gene clusters

- Allows for graphical visualization of distinct evolutionary processes
- Help tree interpretation for branches with low support

Figure 2.3: Schematic diagram of results based on phylogenetic (left) and network (right) analyses. Potential contributions of each approach, as well as benefits stemming from the combination of both methods are described below the diagrams.

tools and statistical approaches developed for phylogenetic analysis, however, it can provide an independent assessment of relationship structures which can aid in the interpretation of phylogenetic data, especially in areas of the tree exhibiting low branch support. In particular, network analysis can be used to generate testable hypotheses regarding the conservation or divergence of gene function in cases of potential subfunctionalization or neofunctionalization. In combination, we believe these methods provide a robust framework that expands the power of gene family evolution studies.
2.6 Application on a C₄ photosynthesis gene family: PEPC genes

One of the enzymes involved in the C₄ photosynthetic cycle is phosphoenol pyruvate carboxylase (PEPC) (Figure 1.1). This enzyme is responsible for catalyzing the first carboxylation step, forming OAA, and is present in all C₄ photosynthesizing subtypes. Here I apply both phylogenetic and network approaches described in this chapter to show how the methods in this chapter can be used in C₄ photosynthesis research, because this enzyme is utilized by all C₄ subtypes.

Methods

Different sorghum PEPC gene copies were obtained from the sorghum genome available at Phytozome version 13, based on the gene copies described in Wang et al. The genomes of Setaria italica and Oryza sativa were also obtained from Phytozome version 13. Zea mays genome version 4 was obtained from maizeGDB. Eragrostis tef and Oropetium thomaeum were obtained from CoGe organism view. The Pennisetum glaucum genome was obtained from http://gigadb.org/dataset/100192. Only the CDS was used for this analysis. These species were selected because their genomes are either well studied or annotated (sorghum, S. italica, Z. mays and O. sativa), or have been released in the past 5 years (E. tef and O. thomaeum).

Orthology of the species studied here was assigned by performing a LASTZ search to find the best reciprocal LASTZ hits to the sorghum PEPC genes. Because both Z. mays and E. tef are tetraploids compared to sorghum, the two best reciprocal LASTZ hits were considered. All the genes retrieved using this approach were present in syntenic regions, except one of the E. tef genes. One of the C₄ PEPC copies of E. tef genes was present in a syntenic position, while the other ortholog was not. This gene, however, presents a second copy in a syntenic position that was not retrieved by the LASTZ search. For this reason, in the downstream analysis both the gene retrieved by the LASTZ search and the
syntenic homologous gene were not retrieved by this approach. The sequence alignment was performed using MUSCLE v3.8.31. The phylogenetic tree was generated using MEGA X using default parameters of the maximum likelihood approach. The remaining steps for network analysis were performed as described in this chapter (see Methods, section 3.2).

**Results**

The phylogenetic analysis successfully retrieved the clusters of the 4 sorghum gene copies analyzed. Similarly, the network approach was able to retrieve the same gene copy clusters. The network analysis showed the peak of 74% as the most distant network topology (Figure 2.4A). The phylogenetic tree clustered the PEPC genes based on the different sorghum gene copies. Similarly, the network generated retrieved the same gene clusters as the phylogenetic tree (Figure 2.4B). The network result provides insights on the fact that the sister group of the PEPC C₄ copy is very different from one another, despite their close evolutionary relationship (Figure 2.4C). Even though the network obtained at the 74% peak retrieved the gene copy clusters, it was already too fragmented to make any extra evolutionary inferences. The last network where all clusters were connected was searched. For this reason, the peak of 72% was analyzed, since this was the last peak where all clusters were connected before being too fragmented to analyze (Figure 2.4B).

Although the phylogenetic tree retrieved the gene copy clusters, it would be challenging to infer which group actually contained the gene copy that evolved the new C₄ related function without prior gene function tests. The network analysis can help overcome this challenge by analyzing both results together, phylogenetic and network results. In the network obtained at 72% it is possible to notice that the genes in the green cluster, sister to the C₄ gene copy cluster in the phylogenetic tree, has a higher number of connections with the other non-C₄ gene copy clusters. Also, the cluster with the C₄ copies
Figure 2.4: Summary of network and phylogenetic results. A) $\delta(\sigma, \sigma+\Delta\sigma)$ distance. The values for the analyzed networks obtained at 72% and 74% are marked; B) Network obtained at 75%; C) Phylogenetic tree obtained with gene copy PEPC gene clusters highlighted with different colors and showing the placement of the $C_4$ PEPC gene copy. Branch length values are shown in the tree; D) Network obtained at 72%. All networks and phylogenetic tree are colored based on the gene copy clusters. Same colors represent the same clusters in the networks and phylogenetic tree. $C_4$ copies are represented in red.

exhibits a weak connection with the green cluster and no connection with the remaining non-$C_4$. This result helps make the case that the red cluster contains the most different copies, and possibly these copies neofunctionalized to perform a $C_4$ function (Figure 2.4D). Finally, these results support the findings of this chapter that using the network approach along with phylogenetic analysis can provide supplementary information about gene family evolution.
2.7 Supplementary information
<table>
<thead>
<tr>
<th>Gene name</th>
<th>Species name</th>
<th>Clade name</th>
<th>ACCESSION or GeneID</th>
</tr>
</thead>
<tbody>
<tr>
<td>GinbiMADS5</td>
<td><em>Ginkgo biloba</em></td>
<td>Ginkgoaceae</td>
<td>GU563899</td>
</tr>
<tr>
<td>PiabdAL2</td>
<td><em>Pinus abies</em></td>
<td>Pinaceae</td>
<td>X79280.1</td>
</tr>
<tr>
<td>PiradAG</td>
<td><em>Pinus radiata</em></td>
<td>Pinaceae</td>
<td>AF023615</td>
</tr>
<tr>
<td>TbachAG</td>
<td><em>Taxus baccata</em></td>
<td>Taxaceae</td>
<td>JF519754</td>
</tr>
<tr>
<td>CryjaMADS4</td>
<td><em>Cryptomeria japonica</em></td>
<td>Taxodiaceae</td>
<td>HM177453</td>
</tr>
</tbody>
</table>

### Basal Angiosperms

<table>
<thead>
<tr>
<th>Gene name</th>
<th>Species name</th>
<th>Clade name</th>
<th>ACCESSION or GeneID</th>
</tr>
</thead>
<tbody>
<tr>
<td>ShenAG</td>
<td><em>Saruma henryi</em></td>
<td>Aristolochiaceae</td>
<td>AY464101</td>
</tr>
<tr>
<td>ChlspiSTK</td>
<td><em>Chloranthus spicatus</em></td>
<td>Chloranthaceae</td>
<td>AY464099</td>
</tr>
<tr>
<td>PeamAG1</td>
<td><em>Persea americana</em></td>
<td>Lauraceae</td>
<td>DQ398021</td>
</tr>
<tr>
<td>PeamAG2</td>
<td><em>Persea americana</em></td>
<td>Lauraceae</td>
<td>DQ398022</td>
</tr>
<tr>
<td>MapreSTK</td>
<td><em>Magnolia figo</em></td>
<td>Magnoliaceae</td>
<td>JQ326236</td>
</tr>
<tr>
<td>MiaAG</td>
<td><em>Michelia alba</em></td>
<td>Magnoliaceae</td>
<td>AB050653</td>
</tr>
</tbody>
</table>

### Monocots

<table>
<thead>
<tr>
<th>Gene name</th>
<th>Species name</th>
<th>Clade name</th>
<th>ACCESSION or GeneID</th>
</tr>
</thead>
<tbody>
<tr>
<td>ElguiAG1</td>
<td><em>Elaeis guineensis</em></td>
<td>Areceae</td>
<td>AY739698</td>
</tr>
<tr>
<td>ElguiAG2</td>
<td><em>Elaeis guineensis</em></td>
<td>Areceae</td>
<td>AY739699</td>
</tr>
<tr>
<td>ElguiSTK</td>
<td><em>Elaeis guineensis</em></td>
<td>Areceae</td>
<td>XP_010912706.1</td>
</tr>
<tr>
<td>BdiAG1*</td>
<td><em>Brachypodium distachyon</em></td>
<td>Poaceae</td>
<td>Bradi2g06330.1</td>
</tr>
<tr>
<td>BdiAG2*</td>
<td><em>Brachypodium distachyon</em></td>
<td>Poaceae</td>
<td>Bradi4g40350.1</td>
</tr>
<tr>
<td>BdiAG3*</td>
<td><em>Brachypodium distachyon</em></td>
<td>Poaceae</td>
<td>Bradi2g25090.1</td>
</tr>
<tr>
<td>OsMADS3</td>
<td><em>Oryza sativa</em></td>
<td>Poaceae</td>
<td>L37528</td>
</tr>
<tr>
<td>OsMADS13</td>
<td><em>Oryza sativa</em></td>
<td>Poaceae</td>
<td>AF151693</td>
</tr>
<tr>
<td>OsMADS21</td>
<td><em>Oryza sativa</em></td>
<td>Poaceae</td>
<td>FJ750944</td>
</tr>
<tr>
<td>OsMADS58</td>
<td><em>Oryza sativa</em></td>
<td>Poaceae</td>
<td>AB232157</td>
</tr>
<tr>
<td>SbAG1*</td>
<td><em>Sorghum bicolor</em></td>
<td>Poaceae</td>
<td>Sb03g002525</td>
</tr>
<tr>
<td>SbAG2*</td>
<td><em>Sorghum bicolor</em></td>
<td>Poaceae</td>
<td>Sb08g006460 or</td>
</tr>
<tr>
<td>SbAG3*</td>
<td><em>Sorghum bicolor</em></td>
<td>Poaceae</td>
<td>Sobic.008G072900.1</td>
</tr>
<tr>
<td>ZAG2*</td>
<td><em>Zea mays</em></td>
<td>Poaceae</td>
<td>GRMZM2G160687</td>
</tr>
<tr>
<td>ZMM2*</td>
<td><em>Zea mays</em></td>
<td>Poaceae</td>
<td>GRMZM2G359952</td>
</tr>
<tr>
<td>ZAG1*</td>
<td><em>Zea mays</em></td>
<td>Poaceae</td>
<td>GRMZM2G052890</td>
</tr>
<tr>
<td>LaschisAG</td>
<td><em>Lacandonia schismatic</em></td>
<td>Triuridaceae</td>
<td>GQ214163</td>
</tr>
<tr>
<td>LaschisSTK</td>
<td><em>Lacandonia schismatic</em></td>
<td>Triuridaceae</td>
<td>GQ214164</td>
</tr>
<tr>
<td>GongaSTK</td>
<td><em>Gongora galeata</em></td>
<td>Orchidaceae</td>
<td>AIU94767.1 or</td>
</tr>
<tr>
<td>GongaAG</td>
<td><em>Gongora galeata</em></td>
<td>Orchidaceae</td>
<td>KF914206.1 or</td>
</tr>
<tr>
<td>HyvilSTK</td>
<td><em>Hypoxis villosa</em></td>
<td>Hypoxidaceae</td>
<td>AIU94766.1 or</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>KF914205.1 or</td>
</tr>
<tr>
<td>Gene name</td>
<td>Species name</td>
<td>Clade name</td>
<td>ACCESSION or GeneID</td>
</tr>
<tr>
<td>-----------</td>
<td>--------------</td>
<td>------------</td>
<td>---------------------</td>
</tr>
<tr>
<td>HyvilAG</td>
<td><em>Hypoxis villosa</em></td>
<td>Hypoxidaceae</td>
<td>AIU94771.1</td>
</tr>
<tr>
<td>AspvirSTK</td>
<td><em>Asparagus virgatus</em></td>
<td>Asparagaceae</td>
<td>AB175825.1</td>
</tr>
<tr>
<td>AspvirAG</td>
<td><em>Asparagus virgatus</em></td>
<td>Asparagaceae</td>
<td>BAD18011.1</td>
</tr>
</tbody>
</table>

### Basal Eudicots

<table>
<thead>
<tr>
<th>Gene name</th>
<th>Species name</th>
<th>Clade name</th>
<th>ACCESSION or GeneID</th>
</tr>
</thead>
<tbody>
<tr>
<td>BgIAG</td>
<td><em>Berberis gilgiana</em></td>
<td>Berberidaceae</td>
<td>AY464106</td>
</tr>
<tr>
<td>EupleAG1</td>
<td><em>Euptelea pleiosperma</em></td>
<td>Eupeleaceae</td>
<td>GU357452</td>
</tr>
<tr>
<td>EupleAG2</td>
<td><em>Euptelea pleiosperma</em></td>
<td>Eupeleaceae</td>
<td>GU357453</td>
</tr>
<tr>
<td>AkquiAG</td>
<td><em>Akebia quinata</em></td>
<td>Lardizabalaceae</td>
<td>AY627635</td>
</tr>
<tr>
<td>AktriAG</td>
<td><em>Akebia trifoliata</em></td>
<td>Lardizabalaceae</td>
<td>AY627629</td>
</tr>
<tr>
<td>HogrAG1</td>
<td><em>Holboelia grandiiflora</em></td>
<td>Lardizabalaceae</td>
<td>JQ806406</td>
</tr>
<tr>
<td>HogrAG2</td>
<td><em>Holboelia grandiiflora</em></td>
<td>Lardizabalaceae</td>
<td>JQ806407</td>
</tr>
<tr>
<td>EscaAG1</td>
<td><em>Eschscholzia californica</em></td>
<td>Papaveraceae</td>
<td>DQ088996</td>
</tr>
<tr>
<td>EscaAG2</td>
<td><em>Eschscholzia californica</em></td>
<td>Papaveraceae</td>
<td>DQ088997</td>
</tr>
<tr>
<td>EscaSTK</td>
<td><em>Eschscholzia californica</em></td>
<td>Papaveraceae</td>
<td>DQ088998</td>
</tr>
<tr>
<td>AqAG1*</td>
<td><em>Aquilegia coerulea</em></td>
<td>Ranunculaceae</td>
<td>Aquca-136-00009.1</td>
</tr>
<tr>
<td>AqAG2*</td>
<td><em>Aquilegia coerulea</em></td>
<td>Ranunculaceae</td>
<td>Aquca-022-00039.1</td>
</tr>
<tr>
<td>AqAGL11*</td>
<td><em>Aquilegia coerulea</em></td>
<td>Ranunculaceae</td>
<td>Aquca-136-00010.1</td>
</tr>
<tr>
<td>ThathAG1</td>
<td><em>Thalictrum thalictroides</em></td>
<td>Ranunculaceae</td>
<td>JN887118</td>
</tr>
<tr>
<td>ThathAG2</td>
<td><em>Thalictrum thalictroides</em></td>
<td>Ranunculaceae</td>
<td>AY867879</td>
</tr>
<tr>
<td>MedilSTK</td>
<td><em>Meliosma dilleniifolia</em></td>
<td>Sabiaceae</td>
<td>AY464105</td>
</tr>
</tbody>
</table>

### Core Eudicots

<table>
<thead>
<tr>
<th>Gene name</th>
<th>Species name</th>
<th>Clade name</th>
<th>ACCESSION or GeneID</th>
</tr>
</thead>
<tbody>
<tr>
<td>AlyrAG*</td>
<td><em>Arabidopsis lyrata</em></td>
<td>Brassicaceae</td>
<td>946287</td>
</tr>
<tr>
<td>AlyrSHP1*</td>
<td><em>Arabidopsis lyrata</em></td>
<td>Brassicaceae</td>
<td>486333</td>
</tr>
<tr>
<td>AlyrSHP2*</td>
<td><em>Arabidopsis lyrata</em></td>
<td>Brassicaceae</td>
<td>321962</td>
</tr>
<tr>
<td>AlyrSTK*</td>
<td><em>Arabidopsis lyrata</em></td>
<td>Brassicaceae</td>
<td>489841</td>
</tr>
<tr>
<td>ATSHP1</td>
<td><em>Arabidopsis thaliana</em></td>
<td>Brassicaceae</td>
<td>AT3G58780</td>
</tr>
<tr>
<td>ATSHP2</td>
<td><em>Arabidopsis thaliana</em></td>
<td>Brassicaceae</td>
<td>AT2G42830</td>
</tr>
<tr>
<td>ATSTK</td>
<td><em>Arabidopsis thaliana</em></td>
<td>Brassicaceae</td>
<td>AT4G09960</td>
</tr>
<tr>
<td>BraAG*</td>
<td><em>Brassica rapa</em></td>
<td>Brassicaceae</td>
<td>Bra.K01743.1</td>
</tr>
<tr>
<td>BraSHP1*</td>
<td><em>Brassica rapa</em></td>
<td>Brassicaceae</td>
<td>Bra.G01817.1</td>
</tr>
<tr>
<td>BraSHP2*</td>
<td><em>Brassica rapa</em></td>
<td>Brassicaceae</td>
<td>Bra.E00310.1</td>
</tr>
<tr>
<td>BraSTK*</td>
<td><em>Brassica rapa</em></td>
<td>Brassicaceae</td>
<td>Bra.C02624.1</td>
</tr>
<tr>
<td>CaruAG*</td>
<td><em>Capsella rubella</em></td>
<td>Brassicaceae</td>
<td>Carubv10005558m</td>
</tr>
<tr>
<td>CaruSHP1*</td>
<td><em>Capsella rubella</em></td>
<td>Brassicaceae</td>
<td>Carubv10019520m</td>
</tr>
<tr>
<td>CaruSHP2*</td>
<td><em>Capsella rubella</em></td>
<td>Brassicaceae</td>
<td>Carubv10025002m</td>
</tr>
<tr>
<td>CaruSTK*</td>
<td><em>Capsella rubella</em></td>
<td>Brassicaceae</td>
<td>Carubv10003771</td>
</tr>
<tr>
<td>ThhSHP1*</td>
<td><em>Thelungiella halophila</em></td>
<td>Brassicaceae</td>
<td>Thhalv10006196</td>
</tr>
<tr>
<td>ThhSHP2*</td>
<td><em>Thelungiella halophila</em></td>
<td>Brassicaceae</td>
<td>Thhalv10017047</td>
</tr>
<tr>
<td>ThhSTK*</td>
<td><em>Thelungiella halophila</em></td>
<td>Brassicaceae</td>
<td>Thhalv10028938</td>
</tr>
<tr>
<td>Species</td>
<td>Genus</td>
<td>Family</td>
<td>Accession Numbers</td>
</tr>
<tr>
<td>----------------------</td>
<td>-----------------</td>
<td>----------</td>
<td>--------------------------------------------------</td>
</tr>
<tr>
<td>CapaSHP*</td>
<td>Carica papaya</td>
<td>Caricace</td>
<td>Evm. TU supercontig_50.73 Medtr8g087860.1</td>
</tr>
<tr>
<td>MetrAG*</td>
<td>Medicago truncatula</td>
<td>Fabaceae</td>
<td>JX308825 Medtr 3g 005530.1</td>
</tr>
<tr>
<td>MetrSHP</td>
<td>Medicago truncatula</td>
<td>Fabaceae</td>
<td>Gorai.N017200.1</td>
</tr>
<tr>
<td>MetrSTK*</td>
<td>Medicago truncatula</td>
<td>Fabaceae</td>
<td>Gorai.01G035500.1</td>
</tr>
<tr>
<td>GoraAG1*</td>
<td>Gossypium raimondii</td>
<td>Malvaceae</td>
<td>Gorai02G042600.1</td>
</tr>
<tr>
<td>GoraAG2*</td>
<td>Gossypium raimondii</td>
<td>Malvaceae</td>
<td>Gorai.009G265100.1</td>
</tr>
<tr>
<td>GoraSHP*</td>
<td>Gossypium raimondii</td>
<td>Malvaceae</td>
<td>Gorai.009G288000.1</td>
</tr>
<tr>
<td>GoraSTK1*</td>
<td>Gossypium raimondii</td>
<td>Malvaceae</td>
<td>Thecc1E6029596t1</td>
</tr>
<tr>
<td>GoraSTK2*</td>
<td>Gossypium raimondii</td>
<td>Malvaceae</td>
<td>Thecc1EG001841t1</td>
</tr>
<tr>
<td>ThecAG*</td>
<td>Theobroma cacao</td>
<td>Malvaceae</td>
<td>Thecc1EG036541t1</td>
</tr>
<tr>
<td>ThecSHP*</td>
<td>Theobroma cacao</td>
<td>Malvaceae</td>
<td>Mgvla012605 or Migut.M00986.1</td>
</tr>
<tr>
<td>ThecSTK*</td>
<td>Theobroma cacao</td>
<td>Malvaceae</td>
<td>Mgvla013047m or Migut.C01334.1</td>
</tr>
<tr>
<td>MguAG*</td>
<td>Mimulus guttatus</td>
<td>Phrymaceae</td>
<td>Potri.01G075800.1</td>
</tr>
<tr>
<td>MguSTK*</td>
<td>Mimulus guttatus</td>
<td>Phrymaceae</td>
<td>Potri.013G04900.1</td>
</tr>
<tr>
<td>PotriAG*</td>
<td>Populus trichocarpa</td>
<td>Salicaceae</td>
<td>Potri.019G077200.1</td>
</tr>
<tr>
<td>PotriSTK*</td>
<td>Populus trichocarpa</td>
<td>Salicaceae</td>
<td>L26295.1</td>
</tr>
<tr>
<td>PotriSTK2*</td>
<td>Populus trichocarpa</td>
<td>Salicaceae</td>
<td>AY098735</td>
</tr>
<tr>
<td>TAG</td>
<td>Solanum lycopersicum</td>
<td>Solanaceae</td>
<td>NM_001247265.2</td>
</tr>
<tr>
<td>TSHP</td>
<td>Solanum lycopersicum</td>
<td>Solanaceae</td>
<td>GSVIVG01000802001</td>
</tr>
<tr>
<td>TSTK</td>
<td>Solanum lycopersicum</td>
<td>Solanaceae</td>
<td>GSVIVT0121303001</td>
</tr>
<tr>
<td>ViviSHP*</td>
<td>Vitis vinifera</td>
<td>Vitaceae</td>
<td></td>
</tr>
<tr>
<td>ViviAG*</td>
<td>Vitis vinifera</td>
<td>Vitaceae</td>
<td></td>
</tr>
</tbody>
</table>

Table 2.1: List of species and sequence identifiers used in this study. Genes retrieved from NCBI (genes with * were retrieved from Phytozome).
<table>
<thead>
<tr>
<th>Symbol</th>
<th>Denomination</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\sigma)</td>
<td>Identity threshold</td>
<td>Threshold value used to build a network, based on similarity values ranging from 0 to 100%. Pairs of sequences that have an identity value greater than or equal to (\sigma) means that they are connected</td>
</tr>
<tr>
<td>(M(\sigma))</td>
<td>Adjacency matrix at (\sigma)</td>
<td>Adjacency matrix obtained at a certain value of (\sigma), composed of 0 and 1, representing whether a pair of sequences is connected (represented by 1) or disconnected (represented by 0)</td>
</tr>
<tr>
<td>(M_{ij})</td>
<td>Element of the adjacency matrix</td>
<td>Represents the presence (1) or absence (0) of an edge between sequences (i) and (j) of an adjacency matrix (M)</td>
</tr>
<tr>
<td>(\hat{M}(\sigma))</td>
<td>Neighborhood matrix</td>
<td>Matrix composed by elements representing the least number of edges necessary to connect a pair of sequences</td>
</tr>
<tr>
<td>(\hat{m}_{ij})</td>
<td>Element of the neighborhood matrix</td>
<td>Represents the least number of edges connecting sequences (i) and (j)</td>
</tr>
<tr>
<td>(\Delta \sigma)</td>
<td>Increments of (\sigma)</td>
<td>Value incremented to (\sigma). i.e. (\Delta \sigma = 1) means that the (\sigma) increases by one</td>
</tr>
<tr>
<td>(\delta(\sigma,\sigma+\Delta \sigma))</td>
<td>Network distance between two networks</td>
<td>Represents the network distance (\delta) between the networks obtained at (\sigma) and (\sigma+\Delta \sigma)</td>
</tr>
</tbody>
</table>

Table 2.2: Summary of symbols
Figure 2.8i: A) Network obtained at 75% identity. Nodes are colored based on the community they belong to in the 67% network, in order to show some community resolution, to highlight the fragmentation of the community structure obtained at 67%. B) Network obtained at 51% identity, also colored based on the community they belong to in the 67% network, despite the fact that the high number of connections did not allow the emergence of community structures. The sequences that do not belong to any community are represented by gray nodes.
Figure 2.S2: A) Network obtained from protein sequences at 60% identity. Nodes are colored based on the same patterns of figures 2.1 and 2.2D. The names on top of the nodes represent the main functional annotation of the clusters obtained.
3: Isoseq transcriptome assembly of new C₃ panicoid grasses provides insights about evolution of Poaceae

3.1 Introduction

The pace of plant genome sequencing has accelerated in recent years. However despite decreases in sequencing costs and improvements in genome assembly quality, species selected for whole genome sequencing often meet one or more of the following criteria: A) agricultural importance, B) status as a genetic model system or C) ecological importance. Sequence data from species which lack direct economic, ecological, or genetic model importance can enable comparative analyses to address biological questions in crops and model species. C₄ photosynthesis has evolved multiple times in the grasses, making it particularly amenable to study through comparative genetic approaches. C₄ photosynthesis requires both substantial biochemical and anatomical changes. All grasses which utilize the C₄ pathway belong to the PACMAD clade, a group of grass subfamilies and tribes which includes substantial numbers of both C₃ and C₄ species. Substantial new insights into both the genes involved in producing the biochemical and anatomical changes required for C₄ photosynthesis, as well as the potential function of individual amino acid residues can be obtained from comparative analysis of individual gene families across species utilizing either C₃ or C₄ photosynthesis within the PACMAD clade. However, assembling sequence data for a single gene family from a large enough set of species through PCR amplification and individual Sanger sequencing remains a time and labor intensive process.

Many domesticated grasses belong to the PACMAD clade, including such as maize
(Zea mays), sugar cane (Saccharum spp.), sorghum (Sorghum bicolor), and foxtail millet (Setaria italica). However every domesticated grass in the PACMAD clade with a sequenced genome utilizes one or more variants of the C₄ photosynthetic pathway.¹²⁴,¹²⁵,¹³⁹,¹⁴⁰ As a result, while published whole genome sequence assemblies exist for at least 14 grasses within the PACMAD clade (Table 1), only one of these (Dichanthelium oligosanthes, a wild species)¹⁴¹ utilizes C₃ photosynthesis. Long-read sequencing can effectively generate sequence for large numbers of full length cDNAs even in species lacking reference genome assemblies.¹⁴²,¹⁴³ One concern with utilizing this technology for comparative genetic studies is that the higher error rate, particularly the frequencies of insertion and deletion errors, make data from long read based sequencing of non-model species unsuitable for use in comparative evolutionary analyses.¹⁴⁴ However, we previously found that observed synonymous substitution rates calculated from consensus sequences constructed using PacBio IsoSeq pipeline were not elevated relative to a sister lineage where gene sequences were taken from a sanger-based whole genome assembly, indicating sequence data obtained in this manner may indeed be suitable for comparative evolutionary analyses.¹⁴⁵

Here we report the sequencing and characterization of IsoSeq based transcriptomes for three additional PACMAD grasses, selecting to enable wider scale studies of protein sequence changes associated with the many parallel origins of C₄ photosynthesis within that clade (Figure 3.1). These species were specifically selected to augment C₃/C₄ comparisons: Hymenachne amplexicaulis, Chasmanthium laxum, and D. oligosanthes. H. amplexicaulis is a member of the grass tribe Paspaleae which contains a mixture of C₃ and C₄ species. The Paspaleae are sister to an exclusively C₄ clade consisting of the two grass tribes Andropogoneae + Arundinelleae which include both maize and sorghum, two species with extensive genomic, genetic, and phenotypic resources. H. amplexicaulis is found in moist habitats and thrives under flooded conditions.¹⁵⁴ Chasmanthium laxum belongs to the grass tribe Chasmanthieae (7 species).
The Chasmanthieae all appear to utilize C₃ photosynthesis and are generally placed as early diverging lineage within the Panicoideae, the grass sub-family containing maize, sorghum, sugar cane, miscanthus, switchgrass, foxtail millet, and proso millet. C. laxum can occur in a variety of environments such as: woods, meadows and swamps. The final species targeted for transcriptome sequences was *Dichanthelium oligosanthes*. *D. oligosanthes* is the only PACMAD species exclusively utilizing C₃ photosynthesis with a published genome sequence to date. It is a member of the grass tribe Paniceae, a group which also includes foxtail millet, proso millet, and switchgrass, but is an outgroup to the MPC C₄ subclade of exclusively C₄-utilizing species within that tribe. As the published *D. oligosanthes* reference genome was constructed utilizing short read sequencing, the inclusion of *D. oligosanthes* provided an opportunity to improve the

<table>
<thead>
<tr>
<th>Species</th>
<th>Relevance</th>
<th>C₃/C₄</th>
<th>Genome Publication</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Dichanthelium oligosanthes</em></td>
<td>Wild Species</td>
<td>C₃</td>
<td>Studer et al. (2016)¹⁴¹</td>
</tr>
<tr>
<td><em>Eleusine coracana</em>³</td>
<td>Grain Crop</td>
<td>C₄</td>
<td>Hittalmani et al. (2017)¹⁴⁶</td>
</tr>
<tr>
<td><em>Eragrostis tef</em>³</td>
<td>Grain Crop</td>
<td>C₄</td>
<td>Cannarozzi et al. (2014),¹⁴⁷</td>
</tr>
<tr>
<td><em>Miscanthus x giganteus</em>³</td>
<td>Biomass Crop</td>
<td>C₄</td>
<td>Swaminathan et al. (2010)¹⁴⁸</td>
</tr>
<tr>
<td><em>Oropetium thomaenum</em>ᵃ</td>
<td>Wild Species</td>
<td>C₄</td>
<td>VanBuren et al. (2015,2018)¹²⁹,¹⁴⁹</td>
</tr>
<tr>
<td><em>Panicum hallii</em>ᶜ</td>
<td>Wild Species</td>
<td>C₄</td>
<td>Lovell et al. (2018)¹⁵⁰</td>
</tr>
<tr>
<td><em>Panicum miliaceum</em>ᶜ</td>
<td>Grain Crop</td>
<td>C₄</td>
<td>Zou et al. (2019)¹⁵¹</td>
</tr>
<tr>
<td><em>Panicum virgatum</em>ᶜ</td>
<td>Biomass Crop</td>
<td>C₄</td>
<td>Casler et al. (2011)¹⁵²</td>
</tr>
<tr>
<td><em>Pennisetum glaucum</em>ᶜ</td>
<td>Grain Crop</td>
<td>C₄</td>
<td>Varshney et al. (2017)¹³⁰</td>
</tr>
<tr>
<td>*Saccharum spp.*ᵇ</td>
<td>Sugar Crop</td>
<td>C₄</td>
<td>Garsmeur et al. (2018)¹⁴⁰</td>
</tr>
<tr>
<td><em>Setaria italica</em>ᵃ</td>
<td>Grain Crop</td>
<td>C₄</td>
<td>Bennetzen et al. (2012)¹²⁵</td>
</tr>
<tr>
<td><em>Setaria viridis</em>ᶜ</td>
<td>Genetic Model</td>
<td>C₄</td>
<td>Brutnell et al. (2010)¹⁵³</td>
</tr>
<tr>
<td><em>Sorghum bicolor</em>ᵇ</td>
<td>Grain/Biomass/Sugar Crop</td>
<td>C₄</td>
<td>Paterson et al. (2009)¹²⁴</td>
</tr>
<tr>
<td><em>Zea mays</em>ᵇ</td>
<td>Genetic Model</td>
<td>C₄</td>
<td>Schnable et al. (2009)¹³⁹</td>
</tr>
</tbody>
</table>

Table 3.1: Published reference genomes for grass species within the PACMAD clade. Species sharing a common inferred evolutionary origin of C₄ photosynthesis as reported in Edwards et al. (2011)²⁰ are indicated by superscript letters.
proportion of genes with full length sequences from this lineage available for comparative analyses. *D. oligosanthes* is present in small glades on the edge of woods (A.J. Studer, personal communication, April 08, 2019). The placement of *C. laxum* as an outgroup to other panicoid grasses with sequenced reference genomes and *D. oligosanthes* as sister to other members of the Paniceae with sequences reference genomes were recovered in a preliminary analysis of our long read dataset. Support of the placement of *H. amplexicaulis* as a sister group to Andropogoneae (sorghum and maize) was strong but not unambiguous.

![Diagram of phylogeny and inflorescences](image)

Figure 3.1: A) Current literature consensus phylogeny of the relationships between the grass species studied here. Lineages in green utilize C₄ photosynthesis, while lineages in black utilize C₃ photosynthesis. The green stars indicate apparent independent origins of C₄ photosynthesis. B) Inflorescence of *H. amplexicaulis*. C) Inflorescence of *C. laxum*. D) Inflorescence of *D. oligosanthes*. 
3.2 Methods

Plant material, RNA extraction, and sequencing

For all three species, young leaf tissue was harvested from mature plants growing in the greenhouses of the University of Nebraska’s Beadle Center, 40.8190, 96.6932, on October 05 2017. Young leaves were harvested from a C. laxum plant germinated from seed collected with accession Kellogg 1268 in Corkwood Conservation Area, just outside of Neelyville, MO, USA. Full details of this collection are published on Tropicos: https://www.tropicos.org/Specimen/100877982. Leaf tissue from D. oligosanthes was harvested from a plant descended from Kellogg 1175, which was collected in Shaw Nature Reserve, west of St. Louis, MO, USA. Full details of this collection are published on Tropicos: http://www.tropicos.org/Specimen/100315254. The specific D. oligosanthes plant used as a tissue donor had experienced at least three generations of selfing relative to the originally collected plant. This selfing occurred via an independent lineage from the F2 plant derived from the same collection which was used to generate the DNA for the D. oligosanthes reference genome. Young leaves were harvested from H. amplexicaulis which had been clonally propagated from collection PH2016. PH2016 was originally collected by Pu Huang in Myakka River state park in Florida, USA on March 22nd, 2016. A clone of this same accession, grown in the same greenhouse, is deposited at the University of Nebraska-Lincoln Herbarium with index number NEB-328848.

Tissue samples were ground in liquid N\textsubscript{2} and then approximately 200mg of powdered tissue was added to 2 μL of TriPure isolation reagent (Roche Life Science, catalog number #11667157001). The RNA samples mixed with TriPure were then separated using chloroform, precipitated using isopropanol, and RNA pellets were washed using 75% ethanol. The samples were air-dried and diluted in RNAsecure (Ambion). Total RNA concentration was measured using a NanoDrop spectrophotometer and the integrity was assessed based on electrophoresis on a 1% agarose gel. 10 μL of total RNA
for each species was shipped to the Duke Center for Genomic and Computational Biology (GCB), Duke University, USA. Concentrations at the time of shipment ranging from 226.07 to 1,374 ng/µL. One IsoSeq library was constructed per species and each library was sequenced using a single SMRT cell on a PacBio Sequel.

**Consensus reads and transcriptome assembly**

Two separate sequence datasets were produced per library: full length (FL) transcripts and non-full length (NFL) transcripts. A given transcript was considered FL if the sequence read contained both 5’ and 3’ adapters as well as poly-A tail and are not redundant to other transcripts. The transcripts lacking the poly-A tail or one of the adapters are instead included in the non-full length dataset. Sequence reads from both files were used to assemble consensus transcriptomes using the software pbtranscript to cluster redundant sequences, part of the SMRT pipe package (version 5.1) with default parameters (https://www.pacb.com/wp-content/uploads/SMRT_Tools_Reference_Guide_v600.pdf).

For each final consensus transcript, the single longest ORF present within that transcript was selected as the CDS sequence for downstream analyses. ORFs were required to include an in frame stop codon but were not required to include an in-frame "ATG" which may result in additional non-translated codons being appended to the 5’ end of the putative CDS, but avoids CDS truncation when the 5’ end of the sequence was not recovered.

**Sequence data set**

CDS file containing only one primary transcript per gene downloaded from Phytozome 12 (https://phytozome.jgi.doe.gov/pz/portal.html) was used from *Brachypodium distachyon*,* Oryza sativa* (rice),* Sorghum bicolor* (sorghum)* and Setaria italic* (foxtail millet).* CDS sequences for version 2 of the *Oropetium thomaeum* (oropetium)*
genome (GenomeID 51527), and the draft *Eragrostis tef* genome (GenomeID 50954) were downloaded from CoGe. CDS sequences for the initial release of the *Pennisetum glaucum* (pearl millet) genome where downloaded from GigaDB. CDS sequences for B73_RefGenV of the *Zea mays* (maize) reference genome was retrieved from Ensembl. In cases where only a complete set of CDS sequences was released for a given species, we arbitrarily selected the longest annotated transcript from a given locus to be the single representative transcript for downstream analyses.

The *Eragrostis tef* genome was also obtained from CoGe under the genome ID: 50954, stored at: https://genomevolution.org/CoGe/OrganismView.pl?oid=38364. CDS from *Pennisetum glaucum* (pearl millet) was obtained from, stored at: http://gigadb.org/dataset/100192 released on 2016-02-29. The latest version of *Eragrostis tef* genome was obtained from The maize genome version 4 was retrieved from MaizeGDB, stored at: ftp://ftp.ensemblgenomes.org/pub/plants/release-36/fasta/zea_mays/cds/. The CDS file from maize contained all coding sequences. For this reason, only the longest isoforms of each maize gene in the CDS fasta file were kept for further analysis. Therefore, each maize CDS was represented by their respective longest isoform. We generated the CDS from the Iso-Seq transcriptomic sequences by retrieving the longest ORF containing a stop codon.

**Putative orthology assignments**

CDS sequences obtained from *H. amplexicaulis, C. laxum* and *D. oligosanthes* as described above were compared to the primary CDS sequences of each annotated gene in the sorghum genome using LASTZ version 1.04.00 with the following parameters: –identity=70 –coverage=50 –ambiguous=iupac, –notransition, and –seed=match12. CDS sequences from the three target species were presumed to belong to an orthologous group as a given sorghum gene if the sorghum CDS sequence and target species CDS
sequence were reciprocally identified as each others high scoring hit in the LASTZ analysis.

Orthologous relationships between sorghum genes and genes in other species with sequenced reference genomes were inferred based on syntenic orthology. For each combination of sorghum and rice, brachypodium, oropetium, teff, foxtail millet, pearl millet, sorghum, and maize all by all LASTZ comparisons were performed using the same parameters described above. The resulting LASTZ output was employed to identify initial syntenic genomic blocks using QuotaAlign with the parameters –tandemNmax=10, cscore=0.5, –merge and –Dm=20. The quota was set to –quota=1:2 for maize and teff, and –quota=1:1 for all other species. Pairwise syntenic block data was merged and polished using the methodology previously described in to obtain the final set of high confidence syntenic ortholog groups employed for all downstream analysis.

Orthology was treated as a transitive property, thus each *H. amplexicaulis*, *C. laxum* or *D. oligosanthes* gene identified as putatively orthologous to a given sorghum gene based on reciprocal best LASTZ hit analysis, was also considered to be putatively orthologous to syntenic orthologs of that sorghum gene identified in each of the other species described above. The final sets of putatively orthologous gene groups including both sequences from published reference genomes and the long read sequencing described here is provided in https://github.com/deCarvalho90/Dissertation_sup_materials/blob/master/Sup_material_1_orthology%20lists.xls.

LASTZ was used to perform all by all comparisons of coding sequence from the primary transcript of each gene from species with published genomes listed in the previous section. Sorghum was used to perform the synteny analyses across all species. Finally the synteny results were merged together based on sorghum genes. The following parameters were used in all LASTZ analyses: –identity=70 –coverage=50 –ambiguous=iupac, –notransition, and –seed=match12. To identify syntenic orthologs between species with assembled reference genomes, LASTZ output was used at the input
for QuotaAlign with the additional parameters --tandemNmax=10, cscore=0.5, --merge and --Dm=20.\textsuperscript{164} The quota settings were --quota=1:1 for the following species: B. distachyon, O. sativa, S. italica, O. thomaeum, P. glaucum. As maize and teff are tetraploids compared to sorghum, the comparisons between sorghum and maize as well as sorghum and teff had a different quota, --quota=1:2 (Z. mays and E. tef). Final syntenic orthologs were assigned based on previous methods.\textsuperscript{165} Orthology was inferred to transcripts of H. amplexicaulis, C. laxum and D. oligosanthes using LASTZ,\textsuperscript{131} checking the best reciprocal hits to sorghum genes. Once orthology was assigned to all species analyzed, only sets of homologous genes with at least one gene copy in each species were included in this analysis. The list of orthologous genes analyzed here can be found in https://github.com/deCarvalho90/Dissertation_sup_materials/blob/master/Sup_material_1_orthology%20lists.xls.

**Sequence alignment, QC, and phylogenetic analysis**

Kalign (v2.04) was used to create a multiple sequence alignment from protein sequences obtained by translating CDS sequences from all genes in a given putatively orthologous gene group. This gapped protein alignment was in turn employed to create a codon-level DNA alignment of the original CDS sequences. GBlocks version 0.91 was run with default parameters to identify high quality portions of the sequence alignment and remove those portions of the alignment not meeting specified quality thresholds.\textsuperscript{166} Alignments including only those portions passing GBlocks filtering were then used as input for RAxML version 8, using the GTRGAMMA model and with a clade of rice and brachypodium specified as an outgroup, to obtain a phylogenetic tree for each group of putatively orthologous genes.\textsuperscript{167} When RAxML was unable to construct a phylogeny in which rice and brachypodium formed monophyletic clade sister to other other taxa the trees were omitted from downstream visualization. To plot all phylogenies, we used Densitree, part of the BEAST2 package, was used to create combined blots of large
numbers of trees. For visualization purposes only, all branches were treated as having equal length in order to improve the ease of visually comparing differences in topology.

As a result of the separate whole genome duplications in the maize and teff lineages, in many cases gene and species trees would contain different numbers of leaf nodes. For gene groups where maize and teff had each fractionated back to single copy status, only a single alignment file was created. If fractionation had already occurred in one lineage, but not the other, two separate alignments were created, each sampling one of the two co-orthologous gene copies from the species with a retained whole genome duplication derived gene pair. When fractionation had not occurred in either lineage, four total alignments were generated per gene group, capturing all possible pairwise combinations of the two teff gene copies and two maize gene copies.

### 3.3 Results and Discussion

The number of raw reads generated per species was largely consistent and ranged from 708,681 to 734,932 (Table 3.2). After clustering both full length and non-full length transcripts to obtain a set of polished consensus transcripts, the number of sequences per species dropped to 164,640 to 193,422 (Table 3.2). The average length of consensus sequences ranged from 925 bp to 1,438 kb (Figure 3.S1). The number of consensus transcripts significantly exceeded the expected number of expressed genes, however, this is consistent with other reference genome-free IsoSeq analyses.\textsuperscript{145, 169, 170} Inflated numbers of consensus transcripts can result from sequencing of multiple alternatively spliced isoforms of the same gene, sequencing of incompletely processed mRNA molecules,\textsuperscript{171} high sequence error rates preventing multiple sequences from the same transcript being collapsed into a consensus, divergent haplotypes of the same locus present in our clonally propagated, wild collected, or partially inbred starting material, or contamination of the original samples with mRNA from non-target organisms.

Raw reads generated for the three species ranged from 708,681 to 734,932. After
clustering both full length non-chimeric and non-full length transcripts to obtain a set of polished consensus transcripts ranging from 164,640 to 193,422 (Table 3.2) and average length ranging from 925 to 1,438 (Figure 3.2). The consensus transcripts were aligned to the sorghum genome to obtain the orthologous genes from the iso-seq data. Those transcripts were aligned to sorghum due to its high quality genome annotation and the fact that the species did not undergo a whole genome duplication. *H. amplexicaulis* had the highest number of consensus transcripts and the lowest number of transcripts aligned to sorghum, while both *D. oligosanthes* and *C. laxum* a similar number of genes orthologous to sorghum genes (Table 3.3).

<table>
<thead>
<tr>
<th>Species</th>
<th>Total reads</th>
<th>Raw data</th>
<th>CCS reads</th>
<th>FL reads</th>
<th>Average FL length</th>
<th>Consensus transcripts</th>
<th>Average consensus transcript length</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Hymenachne amplexicaulis</em></td>
<td>774,952 reads</td>
<td>5.6 GB</td>
<td>772,188 reads</td>
<td>285,037 reads</td>
<td>963 bp</td>
<td>119,142</td>
<td>925 bp</td>
</tr>
<tr>
<td><em>Dichanthelium oligosanthes</em></td>
<td>770,831 reads</td>
<td>10.1 GB</td>
<td>501,862 reads</td>
<td>135,914 reads</td>
<td>1,466 bp</td>
<td>150,334</td>
<td>1,438 bp</td>
</tr>
<tr>
<td><em>Chasmanthium laxum</em></td>
<td>729,770 reads</td>
<td>12.1 GB</td>
<td>649,149 reads</td>
<td>106,566 reads</td>
<td>1,234 bp</td>
<td>164,640</td>
<td>1,234 bp</td>
</tr>
</tbody>
</table>

Table 3.2: Summary statistics for raw and processed long read sequence data generated from each of the three target species.

<table>
<thead>
<tr>
<th>Species</th>
<th>Sorghum gene space coverage</th>
<th>Sorghum syntenic gene space coverage</th>
<th>Transcript alignment rate</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>H. amplexicaulis</em></td>
<td>11,445 genes/34,211 genes (33.5%)</td>
<td>6,402 transcripts/11,800 genes (54.2%)</td>
<td>115,361 transcripts/193,422 transcripts (59.6%)</td>
</tr>
<tr>
<td><em>C. laxum</em></td>
<td>13,446 genes/34,211 genes (39.3%)</td>
<td>7,418 transcripts/11,800 genes (62.8%)</td>
<td>125,357 transcripts/164,640 transcripts (76.1%)</td>
</tr>
<tr>
<td><em>D. oligosanthes</em></td>
<td>14,159 genes/34,211 genes (41.3%)</td>
<td>7,760 transcripts/11,800 genes (65.7%)</td>
<td>171,465 transcripts/190,632 transcripts (89.9%)</td>
</tr>
</tbody>
</table>

Table 3.3: Alignment rates of consensus transcripts generated from each of the three target species to the sorghum gene space.

Alignment of final consensus reads to the sorghum reference genome was employed to estimate coverage of the shared grass gene space for data collected from each of our target species, as well as to assist in further collapsing multiple redundant sequences originating from alternative splicing, incomplete processing, or divergent haplotypes of transcripts originating from a single genetic locus. In all three cases that majority of consensus transcripts could be aligned to known genes in the sorghum genome, with an average of between 9.3 and 12.1 consensus transcripts aligning to each sorghum gene represented in the transcriptome data (Table 3.3). Each of these three target species is predicted to be diploid based on either flow cytometry based estimates of genome size and/or imaging of chromosomes, thus a maximum of two transcripts per locus can be
explained by divergent haplotypes. The high number of consensus sequences aligned per represented sorghum locus suggests that a large proportion of the overall inflation in consensus transcript number from this dataset may result from alternative splice isoforms or sequencing of incompletely processed mRNA molecules. It should also be noted that this analysis will confound lineage specific gene duplications with divergent haplotypes and splice isoforms, however this bias will be consistent across all three species.

For each sorghum gene which aligned to two or more consensus transcripts from the same target species, a single representative transcript was selected for further downstream analysis (See Methods). Between 11,485 and 14,159 sorghum genes had a corresponding representative transcript in a given target species (Table 3.3). Here we were using only single library was constructed per species, rather than multiple libraries constructed using different size fractions, the use of RNA from a single tissue rather than pooled RNA from multiple tissue types, and were conducting comparisons between more distantly related species. However, the total proportion of sorghum genes represented in each transcriptome dataset was not substantially lower than the 14,401 T. dactyloides-maize gene pairs identified in a previous study which implemented all of these best practices.\(^{45}\) This may in part be explained by both sequencing and library preparation improvements between the RSII and Sequel iterations of this sequencing technology.

Manual curation was used to access the coverage and quality of sequences retrieved from these three C\(_3\) photosynthesis-utilizing PACMAD species for five genes known to be involved in C\(_4\) photosynthesis: PPDK, PEPC, NADP-MDH, NAD-ME and DCT2 in C\(_4\) photosynthesis-utilizing PACMAD species. In four cases, the representative transcript identified from each of the three target species spanned every annotated codon in sorghum. The one exception was PPDK where the representative transcript identified for \textit{H. amplexicaulis} lacked the first annotated exon of the annotated gene model in
sorghum (Figure 3.2). Multiple isoforms of the PPDK gene have been described in both maize and sorghum, with the shorter isoform, lacking the same exon absent in *H. amplexicaulis*. This shorter isoform lacks the chloroplast transit peptide and encodes cytosolic PPDK protein not thought to be associated with C4 photosynthesis.

Phylogenetic consistency was assessed using a small subset of genes with high confidence syntenic orthologs identified in species with published reference genomes and representative transcripts identified in each of the three target species. A total of 11,800 genes were identified at syntenic orthologous locations across the genomes of rice, brachypodium, teff, oropetium, pearl millet, foxtail millet, sorghum, and maize. Of these in 2,774 cases no representative transcripts were retrieved from *C. laxum*, *D. oligosanthes*, or *H. amplexicaulis*. These cases likely represent conserved genes that are not expressed in developing photosynthetic tissue. In 1,611 cases, a representative transcript was identified in only one of the three target species, and in 2,276 cases, representative transcripts were identified in two of the three target species. In the remaining 5,139 cases representative transcripts were retrieved for all three target species. The complete lists of each of these sets of conserved syntenic genes and corresponding transcripts from 0, 1, 2, or 3 of the target species is provided as part of Supplemental Material 1 (See Methods).

One potential concern is using transcriptome data from species utilizing C3 photosynthesis to provide sequence data for comparative genetic and evolutionary analyses of C4 is that enzymes involved in the C4 cycle will likely different functions unrelated to photosynthesis in C3 plants, and therefore may not be expressed in photosynthetic tissue and hence be missing from from datasets derived from sequencing cDNAs. Of 31 core C4 genes enumerated in Huang et al. (2016), 20 were part of the set of 11,800 sorghum genes with conserved syntenic orthologs identified in each of the tested grass species with a published reference genome. Hence, these genes are almost certainly present within the genomes of *C. laxum*, *D. oligosanthes*, and *H. amplexicaulis* as well, whether or not they were expressed to sufficient levels to be detected in this
analysis. Of these 20 syntenically-conserved C₄ related genes, sequence data was obtained from all three target C₃ utilizing panicoid species in 16 cases. In the remaining four cases – DCT4c, GLR, NADP-ME and SCL – no putatively orthologous transcript was identified in any of the three species. There were no cases where a syntenically conserved gene linked to C₄ photosynthesis was detected in some, but not all, of the three C₃ utilizing species evaluated.

From the list containing a total of 5,139 conserved orthologous gene groups present in all species 231 were discarded for one of several reasons, listed from most common to least common. 1) In 113 cases the CDS sequence for the O. thomaeum genome included one or more in-frame stop codons. 2) In 61 cases in at least one species represented by isoseq data no stop codon was present in any of the 6 possible open reading frames, indicating either a sequencing error or incomplete 3 prime coverage. 3) In 56 cases a syntenic orthologous gene present in version 2.1 of the B. distachyon genome had been removed or renamed in version 3.1 of the B. distachyon genome. 4) One O. thomaeum de novo predicted gene region was not present in the CDS data.

The remaining set of 4,908 conserved orthologous gene groups were used to generate protein-guided codon multiple sequence alignments (See Methods). A subset of these alignments containing at least 900 nucleotides (300 codons) alignment scored as "high quality" by GBlocks were employed to construct individual gene-level trees (Figure 3.S2). In total 746 trees, representing 275 putatively orthologous gene groups were constructed. Multiple trees resulted from retained duplicate gene pairs resulting from lineage specific whole genome duplications in maize and teff. Each duplication had the potential to create a retained syntenic gene pair which were each co-orthologous to single gene copies in other grass species within the analysis. In order to maintain a consistent number of final nodes, when a retained gene pair was observed in one or both species, multiple sampled trees were generated (See Methods). A modest bias towards over representation of retained – rather than fractionated – genes was observed in the
set of genes which were represented in the transcriptome assemblies from all three target species: 37% (1,843/4,908) of maize genes in this set were retained as duplicate pairs vs 30% of all syntenic maize genes, and 100% of teff genes in this set were retained as duplicate pairs vs 91% of all syntenic teff genes. The rice and brachypodium clade represented a known outgroup as these two species belong to the BEP clade which diverged from the PACMAD clade of grasses early in the evolution of this family. In 46 cases, RAxML was unable to place the rice-brachypodium clade as an outgroup suggested broader issues with orthology assignment, correct ORF identification, or alignment. These trees were not included in downstream analyses.

Figure 3.2: A GEvo panel showing transcript coverage of the C₄ PPDK gene in S. bicolor Sobic.009G132900 in each of the three species tested. Red-brown boxes represent regions of similar sequence identified by BLASTN between the sorghum genome and consensus transcript sequences retrieved from H. amplexicaulis, D. oligosanthes, C. laxum (from top most to bottom most). The bottom track indicates the annotated gene structure, with intronic sequence indicated in gray and exonic sequence indicated in either blue (5’ or 3’ untranslated regions) or green (coding sequence). Top y-axis indicates scale of the displayed genomic region in kilobases.

Among the 700 remaining gene trees, 304 (43%) produced a single topology consistent with the prior literature on the relationship of these species (Figure 3.3). The second and third most common topologies were each represented by less than 7% of all calculated trees, 47 and 44 cases respectively. The second and third most common topologies differed from prior published phylogenies regarding the placement H. amplexicaulis. In the second most common topology H. amplexicaulis was placed sister to all other panicoid grass species other than C. laxum. In the third most common topology H. amplexicaulis was placed sister to the Paniceae. Parallel analysis was conducted using all 4,908 conserved orthologous gene groups, including many cases with substantially
Figure 3.3: Seven hundred distinct phylogenetic trees calculated from separate multiple sequence alignments of 275 putatively orthologous gene groups with large regions of alignment scored as high quality. Blue indicates the most commonly observed topology (304 trees (43% of the total), purple and red indicate the second (47 trees (6.7%) and third most commonly observed topologies (44 trees (6.2%)), respectively.

shorter regions of high quality multiple sequence alignment. The pattern of trees recovered were largely consistent with those in (Figure 3.3). In the "all genes" analysis, the same most common topology was retrieved as in the long alignment only analysis. The second most common topology in the "all genes" analysis corresponds to the most third most common topology in Figure 3.3, while the third most common in the "all genes" topology places C. laxum as sister to the combined Chloridoideae and Panicoideae (Supplemental Figure 3.S3).

3.4 Data availability

The transcriptome data used here for C. laxum, H. amplexicaulis and D. oligosanthes are available at Zenodo under the DOI 10.5281/zenodo.2687865.

3.5 Supplemental figures
Figure 3.S1: Distribution of lengths for polished transcript sequences: A) *H. amplexicaulis*, B) *D. oligosanthes*, C) *C. laxum*.
Figure 3.52: Distribution of alignment lengths after GBlocks cleaning. Dashed line represents the threshold of 900 nucleotides long sequences employed for Figure 3.3. Sequences represented on the right side of the histogram were analyzed.

Figure 3.53: Plot of 10,876 phylogenetic trees. The blue branches represent the most common topology, purple and red branches represent second and third most common topologies, respectively. Figure generated using Densitree.
Reconstructed protein sequence evolution suggests C₄ photosynthesis evolved via a C₂ ancestor in the Paniceae

4.1 Introduction

The C₄ photosynthetic pathway is relevant to world agronomy and food production, since C₄ plants are responsible for over a quarter of global terrestrial photosynthetic productivity,²,³ with C₄ grasses accounting for approximately 18% of global productivity.⁶,⁷ The main reason for the high productivity of C₄ plants is the ability to increase CO₂ ratios around RuBisCO, lowering photorespiration. Photorespiration happens when RuBisCO fixes oxygen instead of CO₂, a process that uses ATP and does not generate any sugars.¹⁴,¹⁷⁴ The ability to increase CO₂ concentrations around RuBisCO confers evolutionary advantages to C₄ plants growing in hot and arid environments. In such conditions C₄ plants are able to keep their stomata closed for longer periods of time, compared to their C₁ counterparts. This advantage to grow in hot and arid conditions is due to better CO₂ use efficiency,¹⁵ which leads to losing less water through evaporation. This mechanism is consistent with the evolutionary rise of C₄ photosynthesis as an adaptation to a drop in CO₂ levels approximately 30 Mya.¹ Also, this C₄ photosynthesis evolved over 60 times independently since its origins.¹⁹

C₄ photosynthesis is generally performed in two different cell types, mesophyll (M) and bundle-sheath (BS) cells. These cells are arranged around the vascular tissue, known as kranz anatomy,¹² although exceptions to this trend have been described.⁵⁸,¹⁷⁵ This evolutionary feature allows CO₂ fixation to happen in two steps, first in the M and then again in the BS cells. In the M cells, CO₂ is fixed into bicarbonate to form
oxaloacetate (OAA). After this first step, OAA is converted to either malate or aspartate. Different plant species may convert OAA to malate, to aspartate or to both, which is then transported to the BS cells to be decarboxylated. The CO₂ generated by the decarboxylation reaction can then be fixed by the Calvin-Benson cycle.\textsuperscript{13}

The C\textsubscript{4} pathway can be classified into three main distinct subtypes depending on the decarboxylase enzyme used.\textsuperscript{15,22} NAD-malic enzyme (NAD-ME), NADP-malic enzyme (NADP-ME) and PEP carboxykinase (PEP-CK), although previous reports have shown plants performing a mix of these subtypes. For instance, maize and foxtail millet have been traditionally classified as performing NADP-ME pathway, and *Panicum hallii* has been classified as performing NAD-ME pathway. However, recent reports indicate that each of these three species express multiple decarboxylase enzymes. Besides the NADP-ME pathway, maize performs an ancillary PEP-CK pathway,\textsuperscript{27,28,176} foxtail millet shows significant expression of NAD-ME in immature tissues and metabolites involved both in NADP-ME and NAD-ME C\textsubscript{4} pathways\textsuperscript{29} and *P. hallii* exhibits significant expression of NADP-ME.\textsuperscript{56} Under greenhouse, natural environment and drought stress conditions, plants performing different C\textsubscript{4} subtype pathways show variances in photosynthetic traits such as: CO₂ fixation and water use efficiency. Previous studies suggest each subtype exhibit differences in photosynthetic traits, such as higher water use efficiency in NAD-ME,\textsuperscript{35} PEP-CK showing intermediate CO₂ fixation rates\textsuperscript{10} and NADP-ME being the most efficient subtype.\textsuperscript{6,36}

The diversity of ways C\textsubscript{4} pathway can be performed leads to questions about the evolution of C\textsubscript{4} subtypes, and their physiological differences. Addressing these questions is challenging due to evolutionary pressures that different C\textsubscript{4} plants have been exposed to in their groups. Among different plant groups performing C\textsubscript{4} photosynthesis, the tribe Paniceae is the only group that encompasses all three C\textsubscript{4} subtype pathways sharing one common ancestor, without C\textsubscript{3} species separating them phylogenetically.\textsuperscript{19,158} Therefore, we focused only on grasses of the Paniceae tribe to investigate the ancestral state of C\textsubscript{4}
Despite the presence of different subtypes, it has been shown that groups of plants that evolved C₄ photosynthesis did not convert back to C₃, but the question about possible ancestral subtypes was not addressed. Even though the phylogeny of the tribe Paniceae has been widely studied, none of these studies address the ancestral photosynthetic state of Paniceae grasses. In addition, studies about ancestral state of the tribe are scarce and do not agree with the same hypothesis to explain the ancestral photosynthetic subtype of Paniceae. Therefore, multiple independent origins of C₄ photosynthesis, along with the fact that the C₄ pathway can be performed in different ways, raises the question of which type, or subtype, of photosynthesis was present in the common ancestor of the Paniceae. The Paniceae tribe contains 84 genera, that includes the subtribes Melinidinae, Panicinae and Cenchrinae (MPC clade).

Previous studies have addressed the evolution of the C₄ pathway in a number of ways. Phylogeny of Paniceae using chloroplast, mitochondrial and nrDNA markers and subtype ancestral state reconstructions are consistent with models where either the ancestor of the tribe employed the NAD-ME subtype of C₄ photosynthesis or employed C₃ photosynthesis and C₄ photosynthesis evolved independently in each lineage utilizing a different decarboxylation enzyme. However, an ancestral state reconstruction using expression data from plants of the MPC clade indicate that the most likely hypothesis for C₄ evolution is that all three subtypes were already present in the MRCA. Another possible approach to this evolutionary question is comparative analysis of protein sequence evolution. Comparisons across species have shown protein changes linked to C₄ photosynthesis evolution. Phylogenetic analysis of the PEPC gene in multiple species has shown that non-C₄ and C₄ copies of the gene exhibit different amino acids in specific protein sites that are known to change catalytic sites and protein folding. Other studies propose that the C₂ photosynthetic cycle is involved in the evolution of the C₄ pathway. Possibly, the main reason for that is the fact that the C₂ cycle uses the
photorespiratory pathway as a CO₂ carbon pump. The transport metabolite of the C₂ cycle is glycine, a two-carbon compound. For this reason this carbon concentrating mechanism using the photorespiratory pathway is called C₂ photosynthesis.¹⁸¹,¹⁸³

Here we use analysis of protein sequence evolution of known C₄ related genes to evaluate four main hypotheses to explain the ancestral state of the most recent common ancestor (MRCA) of the MPC clade: 1) C₃ or non-C₄ common ancestor (Figure 4.1A), 2) NADP-ME subtype was the ancestral state (Figure 4.1B), 3) a mix of both PEPCK and NAD-ME as one subtype was the ancestral state (Figure 4.1C) or 4) all three subtypes were present in the common ancestor of the C₄ plants (Figure 4.1D). The method we use here aims to investigate the evolutionary rates of different enzymes involved in C₄ photosynthesis, and analyze how they behave in their ancestral branch. Our findings suggest that subtype specific enzymes evolve significantly faster than C₃ plants on branches leading to C₄ species rather than the ancestral branch, suggesting these enzymes evolved as a consequence of independent adaptation instead of evolving from a common C₄ common ancestor. Additionally, photorespiratory related and PPDK genes studied here, present in the C₂ pathway,¹⁸¹ show a significantly faster evolutionary rate in the ancestor branch of the Paniceae compared to the C₃ branches. Therefore, our results support a model that the common ancestor of MPC clade C₄ grasses was a C₂ photosynthesis performing plant (intermediate C₂-C₄ plant). This conclusion is consistent with the model in Figure 4.1A, in which the common ancestor of the MPC clade does not utilize the C₄ photosynthetic pathway.

4.2 Material and methods

Plant growth and RNA-Seq data generation for Urochloa fusca

Urochloa fusca seeds were planted and grown in a Percival (Percival model E-41L2) growth chamber with target conditions of 111 μmol m⁻² s⁻¹, 60% relative humidity, a 12 hour/12 hour day night cycle with a target temperature of 29°C during the day and 23°C at night.
Figure 4.1: Models explaining evolution of different subtypes of C₄ photosynthesis. A) C₄ photosynthesis evolved from a C₃/non-C₄ common ancestor and different subtypes evolved independently; B) Ancestor was a C₄ NADP-ME subtype; C) Ancestor was a mix of both C₄ NAD-ME and PEPCK subtypes; D) Ancestor was a C₄ capable of performing all three subtypes. Branches leading to specific subtypes are colored as follows: red - NADP-ME; blue+orange - NAD-ME+PEPCK. Arrows point to the ancestral branches.

Under the growing conditions employed, twelve days old after sowing (DAS) plants were collected and whole seedlings were used for RNA extraction. RNA isolation and library construction followed the protocol described by Zhang et al.¹⁸⁴

**Sequence data set**

Coding Sequences (CDS) for the transcript annotated as "primary" for each gene in *Brachypodium distachyon*,⁴ Oryza sativa (rice),¹²⁶,¹⁶¹ *Panicum hallii¹⁵⁰* and *Setaria italica* (foxtail millet)¹²⁵ were obtained from Phytozome 12.
(https://phytozome.jgi.doe.gov/pz/portal.html). For *Dichanthelium oligosanthes* the CDS were retrieved from version v1.001 in CoGe OrganismView, genome ID 28856 (https://genomevolution.org/CoGe/OrganismView.pl).\textsuperscript{141} CDS from *Pennisetum glaucum* (pearl millet) was obtained from,\textsuperscript{130} stored at http://gigadb.org/dataset/100192 released on 2016-02-29. We obtained *Panicum miliaceum* (proso millet) CDS for the specific genes in this paper from CoGe OrganismView (Genome ID: 52484), version v1.\textsuperscript{151} CDS from *Urochloa fusca* was retrieved by finding the longest ORFs containing a stop codon that were also multiples of 3. Because the 5' end of some sequences was missing we did not look for ORFs with start codons. Sequences for *Urochloa fusca* are available at https://github.com/deCarvalho90/Dissertation_sup_materials/blob/master/Urofus_transc_ALL_CDS.fa.tar.gz.

**Orthology assignment**

LASTZ\textsuperscript{131} was used to perform all by all comparisons of coding sequence from the primary transcript of each gene as downloaded from from Phytozome 12 with the following parameters –identity=70 –coverage=50 –ambiguous=iupac, –notransition, and –seed=match12. LASTZ output was parsed to identify syntenic orthologs using QuotaAlign with the additional parameters –tandemNmax=10, cscore=0.5, –merge and –Dm=20.\textsuperscript{164} To reverse the collapse of tandem gene clusters which are part of the QuotaAlign algorithm, final syntenic orthologs were assigned based on the gene copy with the highest LASTZ alignment score within 20 genes up or downstream of the original syntenic location predicted by quota align. Synteny analysis of proso and pearl millet could not retrieve all $C_4$ related genes here, possibly due to the fact that these genome assemblies are fairly new and might be updated in the future. Due to the reason above, genes that failed to be retrieved by synteny in pearl and proso millet had their orthology inferred by checking the best reciprocal hits to foxtail millet $C_4$ genes using
LASTZ. The dichanthelium genome was generated using short read Illumina technology, which could lead to incomplete sequences. To discard the possibility of syntenic genes not being retrieved because of incomplete sequences, four dichanthelium genes (AK, AspAT, NAD-ME and MEP3b) were retrieved using the same orthology inference used for proso and pearl millet. *Urochloa fusca* orthology was also inferred by reciprocal best LASTZ hits with the foxtail millet C4 genes. The list of C4 genes analyzed here is available at: https://github.com/deCarvalho90/Dissertation_sup_materials/blob/master/Sup_material_C4_genes.xls

**dN/dS calculation and evolutionary analyses**

The CDS of each C4 related gene was translated to protein and then aligned using Kalign version 2.04. A codon-based alignment was created using the protein alignment as a guide. The codon alignment and a guide phylogenetic tree were supplied to the software codeml from PAML package version 4.09 in order to calculate the nonsynonymous and synonymous substitution rates (dN/dS) for each branch of the tree. The guide tree was unrooted and written based on the phylogenetic relationships of the species in this study found in Edwards et al (2011). Finally, the dN/dS values were used to check how fast the ancestral branch leading to the C4 species were evolving and analyze how C4 photosynthesis evolved based on the models in Figure 4.1.

**Statistical comparison of branch dN/dS values and photosynthetic trait values**

A Fisher Exact Test was performed in order to test whether significant differences in evolutionary rate existed between between branches leading to species utilizing C3 photosynthesis and branches leading to species utilizing C4 photosynthesis. The comparison between each one of the C4 branches was performed against the background C3 branches. The C3 background value was calculated by obtaining the average dN/dS ratio of the *O. sativa, B. distachyon* and *D. oligosanthes* branches (excluding their ancestral
branches). The values used to build the contingency tables for the test were obtained from the actual number of synonymous and nonsynonymous substitutions. Because the numbers from the codeml output were not integers, the values were rounded down to be used in the contingency tables. The comparison across subtypes of photosynthesis considering different photosynthetic traits (intercellular carbon dioxide, photosynthetic assimilation and transpiration rates and photosynthetic water use efficiency) was performed using a one-way ANOVA. The same test was used to compare the photosynthetic traits across C₄ photosynthesis independent origins. The one-way ANOVA test was performed including the C₃ species to check if there was a difference between both types of photosynthesis and also just C₄ plants to check if the subtypes and origins are different from one another. Pairwise differences of subtypes in each photosynthetic trait tested were performed with a t-test using python.

**Measurement of photosynthetic assimilation, transpiration, water use efficiency rates and intercellular carbon dioxide**

Photosynthetic and transpiration rates were measured at the Beadle Center greenhouse located on the City Campus of the University of Nebraska-Lincoln, using a portable photosynthesis measurement system (LI-6800 LI-COR, Inc). The species measured were: i) NADP-ME subtype; *Echinochloa colona, Echinochloa esculenta* (japanese millet), *Danthoniopsis dinteri, Setaria italica* (foxtail millet), *Pennisetum glaucum* (pearl millet), *Zea mays* (maize), *Sorghum bicolor* (sorghum) and *Coix lacryma-jobi*, ii) NAD-ME subtype; *Panicum miliaceum* (proso millet) and *Eleusine coracana* (finger millet), iii) PEPCK subtype; *Urochloa fusca* and *Eragrostis tef*. The C₃ relative used here was *Dichanthelium oligosanthes*. These plants were grown for one month in 3 replicates, generating three measurements. Each plant was measured once, representing one replicate. Measurements were performed in 2 adult dichanthelium plants available. Measurements were taken between 9:00 and 14:00 using the first fully expanded leaf that did not present damage of any kind,
under the following LI-COR chamber conditions: 50% relative humidity, at a temperature of 28°C, CO₂ concentration of 400 µmol mol⁻¹, light intensity of 1000 µmol m⁻² s⁻¹.

Because the leaves did not cover the entire LI-COR chamber we measured the leaf width and recalculated the leaf area before each measurement. All measurements were taken once the leaves had stable values for both photosynthetic and stomatal conductance rates. Photosynthetic water use efficiency was calculated using the ratio of photosynthetic and transpiration rates.

4.3 Results

Protein evolutionary rate analysis

Here we investigated 10 genes involved in the C₄ photosynthesis pathway (Figure 4.2) and their evolutionary pattern, based on their subtype classification in the literature and presence of syntenic genes in all species investigated here (Table 4.1). Due to the lack of literature describing NAD-ME and PEP-CK specific genes, we analyzed both NAD-ME and PEP-CK as one C₄ pathway subtype. In this study we considered two different explanations for how C₄ related genes possibly underwent positive selection and change in function (Figure 4.3). One of the explanations is that a C₄ related gene is under positive selection exclusively, which results in a dN/dS ratio greater than 1. In the second explanation such gene could be under different types of selection in its evolution. Therefore the gene could be under purifying selection with a period of positive selection, which results in a dN/dS ratio smaller than 1. The second explanation was the main focus of this study, since we estimated dN/dS ratios smaller than 1 as informative to infer genes involved in C₄ evolution, as long as the values were higher than the values of the same gene in the C₃ branches.

Core C₄ enzymes, those which are utilized by all three C₄ photosynthetic pathways, showed distinct patterns of change in synonymous/nonsynonymous substitution rates relative to enzymes used in specific subtypes. The analysis of the phylogenetic tree was
Figure 4.2: Simplified pathway representation of the three main C₄ photosynthesis subtypes. Enzymes studied here are represented in bold.

<table>
<thead>
<tr>
<th>Enzyme name</th>
<th>C₄ subtype</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>AK</td>
<td>All subtypes</td>
<td>Kanai &amp; Edwards (1999)¹³</td>
</tr>
<tr>
<td>AspAT</td>
<td>NAD-ME &amp; PEPCK</td>
<td>Hatch, Kagawa and Craig (1975), Kanai &amp; Edwards (1999)¹³, ²¹</td>
</tr>
<tr>
<td>DCT2</td>
<td>NADP-ME</td>
<td>Huang et al. (2016)²⁸</td>
</tr>
<tr>
<td>MEP3b</td>
<td>NADP-ME</td>
<td>John et al. (2014)²⁷</td>
</tr>
<tr>
<td>NADP-MDH</td>
<td>NADP-ME &amp; PEPCK</td>
<td>Kanai &amp; Edwards (1999)¹³</td>
</tr>
<tr>
<td>NADP-ME</td>
<td>NADP-ME</td>
<td>Hatch, Kagawa and Craig (1975), Kanai &amp; Edwards (1999)¹³, ²¹</td>
</tr>
<tr>
<td>PEPCK</td>
<td>PEPCK</td>
<td>Hatch, Kagawa and Craig (1975), Kanai &amp; Edwards (1999)¹³, ²¹</td>
</tr>
<tr>
<td>PPDK</td>
<td>All subtypes</td>
<td>Furbank (2011), Hatch, Kagawa and Craig (1975), Kanai &amp; Edwards (1999)¹³, ²¹, ²²</td>
</tr>
<tr>
<td>PPDK-RP</td>
<td>All subtypes</td>
<td>Kanai &amp; Edwards (1999)¹³, ²¹, ²²</td>
</tr>
</tbody>
</table>

Table 4.1: List of enzymes with syntenic orthologs investigated in this study.
Example 1 - C4 under positive selection

\[ C_3: 0 < \frac{dN}{dS} < 1 \]

\[ C_4: \frac{dN}{dS} > 1 \]

Example 2 - C4 under different types of selection

\[ C_3: 0 < \frac{dN}{dS} < 1 \]

\[ C_4: \frac{dN}{dS} > 1 \]

Figure 4.3: Models showing different types of selection and how they affect \( \frac{dN}{dS} \) ratios. Example 1 represents a classical case of positive selection leading to change in function, while Example 2 represents a case where an enzyme might have gone through a mixture of positive and purifying selection leading to a change in function and a final purifying selection period to maintain the enzymatic changes.

PPDK and PPDK-RP, core enzymes, showed a significantly faster evolutionary rates in the common ancestor of C₄ species branch. In contrast another core enzyme, adenylate kinase (AK), only exhibits accelerated protein sequence evolution in certain lineages. AK did not show a fast evolutionary rate in the ancestor branch, but showed a significantly faster evolutionary rate on the proso and pearl millet branches compared to their C₃ counterparts (Figure 4.4).

None of the subtype specific enzymes showed significantly higher \( \frac{dN}{dS} \) values than the background C₃ genes in the common ancestor of C₄ species branch (Figure 4.5). Dicarboxylic acid transporter 2 (DCT2), NADP-ME and MEP3b are enzymes employed in the NADP-ME C₄ pathway. Both NADP-ME and MEP3b enzymes showed a significantly faster evolutionary rate in all branches of the NADP-ME subtype (Foxtail millet, pearl millet and their ancestral branch) and most C₄ branches compared to the background C₃ rate. The C₄ branches where no significant difference was found were: P. hallii in
Figure 4.4: Unrooted phylogenetic trees of $C_4$ photosynthesis core enzymes, present in all subtypes, according to citations in Table 4.1. Branches lengths are equal. Thick red branches represent branches evolving significantly faster than background $C_3$ branches in blue. Abbreviations: Os = Oryza sativa, Bd = Brachypodium distachyon, Do = Dichanthelium oligosanthes, Si = Setaria italica, Pg = Pennisetum glaucum, Uf = Urochloa fusca, Ph = Panicum hallii, Pm = Panicum miliaceum.

NADP-ME enzyme, and both P. hallii and NAD-ME ancestor in MEP3b. DCT2 exhibited a similar pattern to the other two enzymes, with the exception of the branch leading to S. italica. An evolutionary pattern shared among NADP-ME, MEP3b and DCT2 enzymes was a significantly faster evolutionary rate in U. fusca and proso millet, which perform
PEP-CK and NAD-ME subtypes, respectively, compared to the background C₃ branches. NADP-MDH only showed significantly faster evolutionary rates than C₃ branches in pearl millet, NADP-ME species, and the ancestral branch of NAD-ME subtype. Both AspAT and NAD-ME showed significantly higher dN/dS ratios in branches leading to NAD-ME and PEP-CK pathway species compared with C₃ species (Table 4.2).

<table>
<thead>
<tr>
<th>Enzyme name</th>
<th>C₄ subtype</th>
<th>Tree branches</th>
</tr>
</thead>
<tbody>
<tr>
<td>AK</td>
<td>NADP-ME &amp; NAD-ME</td>
<td>pearl millet &amp; proso millet, <em>P. hallii</em> &amp; urochloa*</td>
</tr>
<tr>
<td>AspAT</td>
<td>NAD-ME &amp; PEPCK</td>
<td>proso millet &amp; urochloa</td>
</tr>
<tr>
<td>DCT2</td>
<td>NADP-ME, NAD-ME &amp; PEPCK</td>
<td>ancestral NADP-ME branch, proso millet &amp; urochloa</td>
</tr>
<tr>
<td>MEP3b</td>
<td>NADP-ME, NAD-ME &amp; PEPCK</td>
<td>ancestral C₄ &amp; NADP-ME branches</td>
</tr>
<tr>
<td>NAD-ME</td>
<td>NAD-ME &amp; PEPCK</td>
<td>ancestral NAD-ME branch &amp; urochloa</td>
</tr>
<tr>
<td>NADP-MDH</td>
<td>NADP-ME &amp; NAD-ME</td>
<td>pearl millet &amp; ancestral branch of NAD-ME</td>
</tr>
<tr>
<td>NADP-ME</td>
<td>NADP-ME, NAD-ME &amp; PEPCK</td>
<td>all NADP-ME branches, NAD-ME ancestral branch &amp; urochloa</td>
</tr>
<tr>
<td>PEPCK</td>
<td>PEPCK &amp; NAD-ME</td>
<td>urochloa &amp; proso millet</td>
</tr>
<tr>
<td>PPDK</td>
<td>All subtypes</td>
<td>ancestral branch of all C₄</td>
</tr>
<tr>
<td>PPDK-RP</td>
<td>All subtypes</td>
<td>ancestral branch of all C₄</td>
</tr>
</tbody>
</table>

Table 4.2: List of enzymes with fast evolving tree branches and species. Enzymes in red behaved different than expected according to citations in Table 4.1. Subtypes in bold represent enzymes more likely to belong to that one subtype. *approaching significance, p-value = 0.062.

**Comparison of photosynthetic traits**

As shown in the previous section, differences in evolutionary rates of C₄ related enzymes in different subtypes are an indication of which enzymes are used by each distinct subtype pathway. For instance, the NAD-ME enzyme exhibits a faster evolutionary rate in branches leading to species that perform the NAD-ME and PEP-CK pathways compared to their C₃ counterparts, but such behavior was not seen in the branches leading to NADP-ME species. This example suggests that the NAD-ME enzyme is
Figure 4.5: Examples of C₄ photosynthesis subtype specific enzymes unrooted phylogenetic trees. Branches lengths are equal. Thick red branches represent branches evolving significantly faster than background C₃ branches in blue. Abbreviations: Os = Oryza sativa, Bd = Brachypodium distachyon, Do = Dichanthelium oligosanthes, Si = Setaria italica, Pg = Pennisetum glaucum, Uf = Urochloa fusca, Ph = Panicum hallii, Pm = Panicum miliaceum.

employed in the NAD-ME and PEP-CK pathways, but not employed in the NADP-ME pathway. Differences in enzymes used by each subtype and independent C₄ origins
played a major role in the evolution of the C₄ cycle, which could also shape the photosynthetic yield of plants performing different subtypes and across independent origins. As shown in previous reports,¹⁰,³⁴,³⁵ plants using different C₄ subtype pathways vary in photosynthetic physiological rates.

The t-tests performed to assess differences across C₄ plants classified in the three subtypes showed significant differences between C₃ values and each one of the subtypes for intercellular carbon dioxide (Figure 4.6A), photosynthetic assimilation rate (Figure 4.6B) and photosynthetic water use efficiency (Figure 4.6D) (p < 0.01, t-Test). However no significant difference between the C₃ and C₄ plants was detected for transpiration rate (p > 0.05, t-Test) (Figure 4.6C). Although all subtypes were significantly different from C₃ values for the first three traits, the pairwise comparison of the subtypes did not show a significant difference among them in any of the traits analyzed here (Figure 4.6) (p > 0.05, t-Test).

Figure 4.6: Photosynthetic parameters measurements for C₄ photosynthesis subtypes based on: A) intercellular carbon dioxide; B) photosynthetic assimilation rate; C) Transpiration rate; D) photosynthetic water use efficiency. P-value of pairwise comparisons < 0.05.
We also analyzed how plants utilizing different C₄ subtype pathways behaved across 5 independent origins of C₄ photosynthesis, as the differences among them could be masked by the fact that we analyzed subtype traits together not considering their independent origin. A significant difference was observed when analyzing the intercellular carbon dioxide (Figure 4.7A), photosynthetic assimilation (Figure 4.7B), transpiration (Figure 4.7C), and water use efficiency rates (Figure 4.7D) (p < 0.01, ANOVA). Similarly, a significant difference was observed across C₄ photosynthesis independent origins when analyzing only C₄ plants (p < 0.05, ANOVA).

Figure 4.7: Photosynthetic parameters measurements for five different origins of C₄ photosynthesis based on: A) intercellular carbon dioxide; B) photosynthetic assimilation rate; C) Transpiration rate; D) photosynthetic water use efficiency. Numbers 1-5 represent independent origins of C₄ photosynthesis represented by the following species: 1) *Eleusine coracana*, *Eragrostis tef*; 2) *Danthoniopsis dinteri*; 3) *Echinochloa esculenta*, *Echinochloa colona*; 4) *Panicum miliaceum*, *Pennisetum glaucum*, *Setaria italica*, *Urochloa fusca*; 5) *Zea mays*, *Sorghum bicolor*, *Coix lacryma-jobi*. The colors represent different subtypes present in a certain origin: green (NAD-ME); pink (NADP-ME); blue (PEP-CK). Boxes with multiple colors represent origins with more than one subtype.
Evolution of photorespiratory genes

In addition to the 10 C₄ photosynthesis related enzymes studied here, the evolutionary rate calculation approach used to find genes involved in the C₄ cycle retrieved two enzymes involved in the photorespiratory pathway: glycolate oxidase (GOX) and serine hydroxymethyltransferase (SHMT). They exhibit faster evolutionary rates in C₄ branches compared to background C₃ rates. GOX is localized in the peroxisome while SHMT is localized in the mitochondria (Figure 4.8). While GOX showed an evolutionary pattern similar to both PPDK and PPDK-RP, core enzymes, SHMT behaved similar to subtype specific enzymes. GOX showed higher dN/dS ratio in the C₄ common ancestor branch compared to C₃ counterparts. Also, this enzyme has increased dN/dS ratios in the branches leading to PEP-CK, the ancestral NADP-ME branch and S. italica. SHMT showed faster evolutionary rates in the branches leading to PEP-CK and proso millet (NAD-ME) (Figure 4.9).

4.4 Discussion

In this study we proposed four models to explain the evolution of C₄ photosynthesis and its different subtypes (Figure 4.1): i) common C₃/non-C₄ ancestor (Figure 4.1A); ii) C₄ ancestor performing NADP-ME subtype (Figure 4.1B); iii) performing the NAD-ME+PEP-CK subtype (Figure 4.1C); and iv) performing all subtypes (Figure 4.1D). Protein sequence evolutionary rate data from core C₄ enzymes are consistent with a C₄ common ancestor for the MPC clade within the Paniceae. However, data from subtype specific enzymes are consistent with a non-C₄ common ancestor, possibly a C₃-C₄ intermediate common ancestor performing C₂ photosynthesis, which means that there was subsequent independent evolution of the C₄ subtypes in the MPC clade. The latter scenario proposing a non-C₄ ancestor is consistent with one of the models proposed based on ancestral state reconstruction in Washburn et al (2015). Also, previous
reports showed that C$_3$-C$_4$ intermediate plants are part of the path to explain the evolution of the C$_4$ pathway.$^{189}$

Here we explore the possible explanations for evolution of the C$_4$ subtype pathways comparing protein evolutionary rates of known C$_4$ enzymes with their C$_3$ counterparts. DCT2 enzyme showed faster evolutionary rate leading to the NADP-ME enzyme ancestral branch, indicating that this enzyme is mainly used in the NADP-ME pathway, although DCT2 showed significantly higher dN/dS ratios in NAD-ME or PEP-CK subtypes. It has been shown that species from both NAD-ME and PEP-CK pathways exhibit low expression levels of NADP-ME pathway enzymes, such as NADP-MDH and NADP-ME,$^{56}$ even though DCT2 was not previously described as part of the NAD-ME or PEP-CK pathways. Therefore, we hypothesize that both NAD-ME and PEP-CK pathways could have low levels of malate transport into the BS chloroplast, although we do not have enough data to either support or refute this hypothesis. Despite the fact MEP3b showed signs of fast evolutionary rate in proso millet and \textit{U. fusca}, classified as NAD-ME and PEP-CK pathways, respectively, MEP3b shows significantly faster evolution in all NADP-ME branches compared to the C$_3$ branches. This suggests that MEP3b seems to be mainly used in the NADP-ME subtype. Although MEP3b was not previously described as part of the NAD-ME and PEP-CK pathawys, the fast evolutionary rate of this enzyme is consistent with the aspartate and alanine cycling between M and BS cells present in both subtype pathways.$^{13,22}$ In the aspartate and alanine cycling alanine is converted to pyruvate and transported into M chloroplast. Therefore we suggest pyruvate transport is being performed through MEP3b in the NAD-ME and PEP-CK subtypes.

The NADP-ME tree (Figure 4.5) shows a significantly faster evolutionary rate relative to the C$_3$ counterparts in all NADP-ME branches, PEP-CK branch, the ancestral NAD-ME branch and one of the proso millet genes. Our results showed fast evolving NADP-ME enzyme in the ancestral branch of the NAD-ME subtype species, which supports the possibility of \textit{P. hallii} using both NAD-ME and NADP-ME decarboxylating
enzymes in the photosynthetic pathway, consistent with some level of NADP-ME expression described in *P. hallii*. The NADP-MDH enzyme showed significantly faster evolutionary rate in one NADP-ME subtype species, pearl millet, and in the ancestral branch of NAD-ME species compared with C₃ branches. This is in disagreement with the literature (Table 4.1). However this gene is annotated as "lactate/malate dehydrogenase" in the foxtail millet genome, even though it has been previously annotated as NADP-MDH. Therefore, it is possible that the MDH gene is actually NAD-MDH, enzyme used in the NAD-ME pathway, which could possibly explain why we notice a fast evolutionary rate in the ancestral branch of the NAD-ME species.

Here we combined both NAD-ME and PEP-CK as one subtype due to the lack of literature describing enzymes exclusive to either subtypes. AspAT and NAD-ME enzymes were the only cases where fast evolutionary rates were found in branches leading to the expected subtypes NAD-ME and PEP-CK pathways (Table 4.2). NAD-ME enzyme shows a significantly faster evolutionary rate than background C₃ branches in species performing NAD-ME and PEP-CK subtypes. However, PEP-CK enzyme only shows such evolutionary pattern in *U. fusca* (PEP-CK subtype) and proso millet (NAD-ME subtype). Also, *U. fusca* and proso millet, classified as PEP-CK and NAD-ME subtype, respectively, are the only species with a fast evolving PEP-CK, suggesting that either: 1) proso millet performs some level of the PEP-CK pathway, or 2) that PEP-CK and NAD-ME are, indeed, one subtype. Expression data supports scenario 1 showing that PEP-CK is the only highly expressed decarboxylating enzyme in *U. fusca*, and supporting that such subtype can be performed without any additional supplementary pathways. The scenario 2 seems less likely because that would indicate that both *U. fusca* and proso millet evolved the NAD-ME+PEP-CK pathway independently while *P. hallii* only performed the NAD-ME subtype or that the PEP-CK pathway was lost in *P. hallii*. Either possibility for scenario 2 show problems with their explanations. For instance, independent evolution of the NAD-ME+PEP-CK pathway is not a evolutionary parsimonious explanation, since it
would require the pathway evolved independently twice. Also, to explain the loss of the PEP-CK pathway in *P. hallii* we would expect our results to show a faster evolutionary rate in the ancestral branch of both *Panicum* species along with the fast evolving proso millet branch. Such evolutionary pattern was not observed in our results, which does not support the explanation provided for the loss of the PEP-CK pathway in *P. hallii*.

In addition to the presence of distinct subtypes based on evolutionary analyses, here we analyzed photosynthetic differences across subtypes. Grasses performing NADP-ME subtype have the highest average CO₂ uptake yield under similar conditions.⁶,³⁶ We expected that such subtype would show that photosynthetic trait values are significantly different from the other subtypes (Figure 4.6). However, no significant difference was detected among subtypes. Our results suggest that C₄ plants from different subtypes grown under well watered greenhouse conditions have similar performances, which is in disagreement with previous findings.⁶,³⁶ These studies are subjected to confounding variables such as different plant growth rates, environmental and geographical origins, which could be debatable. Also, we performed our analysis setting the leaf chamber temperature to 28°C, which might have not been a high enough temperature to notice any differences. Independent origins of C₄ photosynthesis show significant differences across different origins (Figure 4.7). Within the Paniceae tribe there is also a significant difference among some species and subtypes (Figure 4.10). Differences between independent origins vary across a wide range, which could be affecting the power of our analysis to detect variation among different subtypes.

Photorespiration has been suggested to play a role in the evolution of C₄ photosynthesis acting as a minor carbon concentrating mechanism around RuBisCO.³ In this study we found two enzymes involved in the photorespiratory pathway that could be involved in the evolution of the C₄ pathway. GOX exhibits the same evolutionary pattern as a core enzyme, while SHMT behaved as a subtype specific enzyme (Figure 4.9). The fact that GOX showed a fast evolutionary rate in the common ancestor of the MPC
clade is consistent with the use of the C\textsubscript{2} photosynthetic pathway present in C\textsubscript{3}-C\textsubscript{4} intermediate plants,\textsuperscript{190} which supports the hypothesis that photorespiration was a bridge between C\textsubscript{3} and C\textsubscript{4} photosynthesis evolution.\textsuperscript{16, 59, 182}

Here we propose that the model suggesting C\textsubscript{4} photosynthesis evolved via a non-C\textsubscript{4}, possibly an intermediate C\textsubscript{3}-C\textsubscript{4} common ancestor performing C\textsubscript{2} photosynthesis and equally predisposed to evolve all three subtypes of photosynthesis is the most likely to explain the evolution of C\textsubscript{4} photosynthesis in the Paniceae (Figure 4.1A). This model is consistent with one of the hypotheses described in a previous report,\textsuperscript{158} where the subtypes could have evolved from multiple C\textsubscript{3} ancestors (possibly C\textsubscript{3}-C\textsubscript{4} predisposed to C\textsubscript{4}). Another possible scenario is that the same amino acid changes were selected for in multiple parallel lineages. If this is the case, in the phylogenetic analysis these changes would be assigned to the common ancestor of those lineages. When calculating dN/dS ratios for the ancestral branch the amino acids assigned to the common ancestor would not be considered as under positive selection, as they would also be present in the parallel lineages. This scenario would mask the positive selection process involved in the evolution of the common ancestor protein sequence, which could explain the fact that AK (core enzyme) did not show a fast evolutionary rate in the common ancestor branch. This may mask a possible C\textsubscript{4} common ancestor for the different subtype pathways. Such outcome is expected as reported previously,\textsuperscript{32, 180} showing that phylogenetic trees using all three codon positions of PEPC and PEPCK genes, results in C\textsubscript{4} photosynthesis evolving from one common ancestor, even though the C\textsubscript{4} pathway evolved independently approximately 60 times.\textsuperscript{20}

In summary, our study showed that core and subtype specific enzymes provide different insights on the possible ancestral state of the common ancestor of the C\textsubscript{4} plants of the Paniceae tribe. While core enzymes suggest a C\textsubscript{4} common ancestor for the Paniceae, subtype specific enzymes suggest a C\textsubscript{3} ancestor. Two core enzymes (PPDK and PPDK-RP) and photorespiratory genes (GOX and SHMT) analyzed here are used in the
$C_2$ photosynthesis cycle support the model suggesting that the ancestral state of the common ancestor of the Paniceae was a non-$C_4$ relative. We suggest this ancestor was most likely an intermediate $C_3$-$C_4$ plant performing $C_2$ photosynthesis. Therefore, we support the model in Figure 4.1A, which suggests a non-$C_4$ common ancestor for the Paniceae.
Figure 4.8: Simplified pathway representation of the three main C₄ photosynthesis subtypes including the C₃, C₂, and photorespiratory pathways. Enzymes studied here are represented in bold. Mitochondrial pathway of the C₂ cycle is the same as the mitochondrial photorespiratory cycle. However the mitochondrial pathway occurs in the bundle sheath cell in the C₂ cycle and in the mesophyll cell in the photorespiratory cycle. Adapted from ²², ²⁸, ¹⁸¹, ¹⁸⁸.
Figure 4.9: Unrooted phylogenetic trees of C₄ photosynthesis photorespiratory enzymes. Branches lengths are equal. Thick red branches represent branches evolving significantly faster than background C₃ branches in blue. Abbreviations: Os = Oryza sativa, Bd = Brachypodium distachyon, Do = Dichanthelium oligosanthes, Si = Setaria italica, Pg = Pennisetum glaucum, Uf = Urochloa fusca, Ph = Panicum hallii, Pm = Panicum miliaceum.
Figure 4.10: LI-COR results for the species in the Paniceae tribe. A) Intercellular carbon; B) photosynthetic assimilation rate; C) transpiration rate; D) photosynthetic water use efficiency. Pm = Panicum miliaceum, Pg = Pennisetum glaucum, Si = Setaria italica, Uf = Urochloa fusca. Boxes are colored by subtype: green, NAD-ME; pink, NADP-ME; blue, PEPCK. P-value of pairwise comparisons < 0.05.
References


[48] Sarah Covshoff, Marek Szecowka, Thomas E Hughes, Richard Smith-Unna, Steven Kelly, Karen J Bailey, Tammy L Sage, Justin A Pachebat, Richard Leegood, and


[56] Jacob D Washburn, Satya S Kothapalli, Julia M Brose, Sarah Covshoff, Julian M Hibberd, Gavin C Conant, and J Chris Pires. Ancestral reconstruction and c3 bundle sheath transcript abundance in the paniceae grasses indicate the foundations for all three biochemical c4 sub-types were likely present in the most recent ancestor. *bioRxiv*, page 162644, 2017.


[179] Osvaldo Morrone, Lone Aagesen, Maria A Scataglini, Diego L Salariato, Silvia S Denham, Maria A Chemisquy, Silvana M Sede, Liliana M Giussani, Elizabeth A Kellogg, and Fernando O Zuloaga. Phylogeny of the paniceae (poaceae:}


[183]


Appendix
A: High-throughput imaging and phenotyping dataset of C₄ grain crops

A.1 Introduction

Studies predict that by 2050 the world population crop production demands can only be met if we are able to double the current production. After the relevant increase in yield after the green revolution, yield production for major food crops as rice, maize and wheat are plateauing. Despite the efforts in both the public and private sectors to develop crops with higher yield, there is a constant need for new approaches to improve plant breeding. Increasing productivity of minor crops neglected by breeding efforts due to the fact they are only grown in specific regions and not worldwide could benefit food production and security. Such crops are referred to as orphan crops. Some of those orphan crops can grow in marginal lands not used for planting major crops and can grow in extreme environments unsuitable for crops such as maize and rice. Among the orphan crops are: finger millet, japanese millet, proso millet, pearl millet, foxtail millet and teff. These crops have not been targeted by breeding efforts and possibly have not been included in phenotyping studies.

Teff and the millets mentioned previously are crops grown mostly in Africa and Asia in regions that are either too hot/arid or too cold for growing major food crops. Teff is a very important food crop in Ethiopia, where its grain is harvested to make flour. The nutritional value of teff is similar to other food crops, but has higher calcium and iron content compared to crops like barley, wheat and rice. Besides its nutritional value, teff can grow in water limited environments as well as poor soils. Finger, pearl and proso millet are food crops that grow in Africa and Asia. These crops can also grow in dry
regions and marginal land. Foxtail millet is one of the oldest crops in the world. It is an important crop in China, which is used for food as well. This millet grows in warm and temperate areas, as well as poor or marginal soils.\textsuperscript{195–197} As well as other millets, japanese millet is a food crop in Africa and Asia. This crop has been studied as an alternative food crop and to learn more about its nutritional value\textsuperscript{198} and as a cover crop.\textsuperscript{199}

Orphan crops can serve as alternatives to increase food production worldwide, hence, increasing the phenotyping efforts of such species could help breeders improve their production. Image data generated from phenotyping can help the identification of agronomic traits important for breeding such as plant height, biomass and yield. The identification of agronomically relevant traits can be achieved in less time using high-throughput phenotyping (HTP). Application of HTP is becoming more common for studying these plant traits.\textsuperscript{200–203} Although HTP approaches are being more used, wheat, barley and rice are among the crops most studied.\textsuperscript{204} Possibly the reason for that is the plant stature of taller plants such as maize and sorghum. Capturing the entire life cycle of plants allows analysis of growth rate as well as how other traits change along plant development. The University of Nebraska-Lincoln Innovation Greenhouse LemnaTec phenotyping system supports imaging plants as tall as \textasciitilde2.5 meters, which allowed us to collect images for 4 months and capture the entire life cycle of the crops grown.

Being able to capture the complete life cycle of plants can be very useful for the development of plant crop growth models. Such models are important for predicting how a plant responds to environmental changes and how it affects its yield production. Several software to model crop growth and make predictions are available.\textsuperscript{205} Images from plant growth from seedling stage all the way to maturity can be used to calibrate those models serving as ground truth data for all different stages of plant development. Imaging plants for their complete life cycle allows the analysis of traits such as plant height, biomass and water use efficiency over time and developmental stage. Also, studies showing the correlation of images and such traits have been performed.\textsuperscript{203, 206}
Image data paired with gas exchange rates are scarce. Ground truth data from gas exchange measurements can help to improve models predicting photosynthetic traits and activity. Here we generated complete life cycle images from RGB, hyperspectral, thermal infrared and fluorescent cameras for 7 C₄ grain crop species that include different genotypes of maize and sorghum (Table A.1). We also collected images from sorghum genotype BTx623 paired with gas exchange rates from a portable photosynthesis measurement system (LI-6800, LI-COR, Inc).

A.2 Methods

Experimental design

Plants were grown and imaged in the greenhouse facility of the University of Nebraska-Lincoln’s Greenhouse Innovation Center. The experiment comprising Eleusine coracana, Eragrostis tef, Echinochloa esculenta, Panicum miliaceum, Pennisetum glaucum, Setaria italica, Sorghum bicolor and Zea mays (Table A.1) was conducted between 13 October 2017 to 5 February 2018, from now on referred to as "experiment A". Another experiment comprising only sorghum plants was conducted between 11 January 2018 to 30 January 2018, from now on referred to as "experiment B". The photoperiod was 12 hours, from 7:00 to 19:00 each day, with supplemental light-emitting diode (LED) growth lamps turned on during that period of time. LED lamps were on during the entire photoperiod as the building infrastructure could block some of the natural light. Temperature in the greenhouse was set to 28°C during the day and 21°C during the night. The soil mix used was Pro-mix 2.8 cubic feet, containing 666 grams of Lime, 10.5 oz (297.67g) of 3-4 month release Osmocote and 10.5 oz (297.67g) of 5-6 month release Osmocote.
Experiment A

Seeds were planted in 10 October 2017 in 1.5 gallon pots containing soil mix, with dry pot weight of 4300 grams when seeds were planted. The pots were weighed every other day and watered back to a target weight. Throughout the experiment the plants were subjected to different target weights. The starting target weight was set to 7200g between October 11 and November 4, the target weight was lowered to 7000g between November 5 and January 10. The target weight was lowered considering how much water the plants were using and visual analysis of the soil to check if it was too wet. As the plants grew the target weight was increased back to 7200g between January 11 and February 5 to meet the water need of plants as they got taller.

All seeds were planted October 10th 2017, but *Setaria italica* genotype N-Si-2 did not germinate in the pot. For this reason, seeds for this genotype were replanted October 17th 2017, one week after the start of the experiment. Three pots with soil, without plants, were imaged with the 15 pots as a control for the amount of water lost through nontranspiration mechanisms (e.g. evaporation). Empty control pots were placed separating plants in groups of 5. All pots were randomized to make sure plants from different species, genotypes, locations, C₄ subtypes and origins were randomly assigned to a position in the LemnaTec belt, with 0.235-meter spacing between plants. The experiment was performed in one row that fit all 18 pots imaged (Table SA.1).

Experiment B

Seeds from sorghum Btx623 were planted in 01 December 2017 in 1.5 gallon pots containing the same soil mix as experiment A, with pot weight of 4300 grams when seeds were planted. Plants were grown for 42 days old before start of imaging. During that time plants were kept off the conveyor belt and moved to the belt January 10th. As in experiment A, pots were weighed and watered back to a target weight of 6800g in January 11th and 12th. Because the soil was still wet a few days after watering the target
<table>
<thead>
<tr>
<th>Scientific name</th>
<th>Common name</th>
<th>Genotype</th>
<th>Geographical Origin</th>
<th>(C_4) subtype</th>
<th>(C_4) origin</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eleusine coracana</td>
<td>Finger millet</td>
<td>I. E. 595</td>
<td>India</td>
<td>NAD-ME</td>
<td>1</td>
</tr>
<tr>
<td>Eragrostis tef</td>
<td>Teff</td>
<td>R286</td>
<td>Ethiopia</td>
<td>NAD-ME</td>
<td>1</td>
</tr>
<tr>
<td>Eragrostis tef</td>
<td>Teff</td>
<td>Red DABI</td>
<td>Ethiopia</td>
<td>NAD-ME</td>
<td>1</td>
</tr>
<tr>
<td>Echinochloa esculenta</td>
<td>Japanese millet</td>
<td>BN-14698-63</td>
<td>United States, Florida</td>
<td>NADP-ME</td>
<td>2</td>
</tr>
<tr>
<td>Panicum miliaceum</td>
<td>Proso millet</td>
<td>Earlybird</td>
<td>United States, Nebraska</td>
<td>NAD-ME</td>
<td>3</td>
</tr>
<tr>
<td>Pennisetum glaucum</td>
<td>Pearl millet</td>
<td>CO42</td>
<td>Niger</td>
<td>NADP-ME</td>
<td>3</td>
</tr>
<tr>
<td>Pennisetum glaucum</td>
<td>Pearl millet</td>
<td>NM-6R1</td>
<td>United States, Nebraska</td>
<td>NADP-ME</td>
<td>3</td>
</tr>
<tr>
<td>Setaria italica</td>
<td>Foxtail millet</td>
<td>Yugu1</td>
<td>China</td>
<td>NADP-ME</td>
<td>3</td>
</tr>
<tr>
<td>Setaria italica</td>
<td>Foxtail millet</td>
<td>N-Si-2</td>
<td>United States, Nebraska</td>
<td>NADP-ME</td>
<td>3</td>
</tr>
<tr>
<td>Sorghum bicolor</td>
<td>Sorghum</td>
<td>BTx623</td>
<td>United States, Texas</td>
<td>NADP-ME</td>
<td>4</td>
</tr>
<tr>
<td>Sorghum bicolor</td>
<td>Sorghum</td>
<td>CM291</td>
<td>United States, Nebraska</td>
<td>NADP-ME</td>
<td>4</td>
</tr>
<tr>
<td>Sorghum bicolor</td>
<td>Sorghum</td>
<td>CM052</td>
<td>United States, Nebraska</td>
<td>NADP-ME</td>
<td>4</td>
</tr>
<tr>
<td>Zea mays</td>
<td>Maize/Corn</td>
<td>B73</td>
<td>United States, Iowa</td>
<td>NADP-ME</td>
<td>4</td>
</tr>
<tr>
<td>Zea mays</td>
<td>Maize/Corn</td>
<td>Va99</td>
<td>United States, Virginia</td>
<td>NADP-ME</td>
<td>4</td>
</tr>
<tr>
<td>Zea mays</td>
<td>Maize/Corn</td>
<td>DE811</td>
<td>United States, Delaware</td>
<td>NADP-ME</td>
<td>4</td>
</tr>
</tbody>
</table>

Table A.1: Summary of species information. *plant was one week younger than the rest of the plants in the experiment.

weight was lowered in different days throughout the experiment. The target weight was 6700g between 13 and 15 January, 6400g 16 January, 6100g between 17 and 19 January, and 5800g until 30 January. In this experiment three sorghum plants were grown and imaged along with two control pots (with soil, without plants).
Measurement of photosynthetic assimilation, transpiration, water use efficiency rates and intercellular carbon dioxide

Photosynthetic and transpiration rates were measured for sorghum plants in experiment B at the greenhouse facility of the University of Nebraska-Lincoln Greenhouse Innovation Center, using a portable photosynthesis measurement system (LI-6800 LI-COR, Inc). The plants were measured on the first day of imaging (12 January 2018, day 1), and on days 4, 6, 8, 10 and 12 of the experiment. Measurements were taken between 9:00 and 14:00 using the first fully expanded leaf that did not present damage of any kind, under the following LI-COR chamber conditions: 50% relative humidity, at a temperature of 28°C, CO₂ concentration of 400 µmol mol⁻¹, light intensity of 1000 µmol m⁻² s⁻¹. All measurements were taken once the leaves had stable values for both photosynthetic and stomatal conductance rates.

Plant imaging

The plants were imaged every other day using four types of cameras, where each camera imaged plants in separate chambers. Images were taken from the different cameras in the following order: 1) thermal infrared; 2) fluorescence; 3) conventional RGB; and 4) hyperspectral. Sequential imaging of the plants was done starting from position 01 and concluding with position 18 (Table 2S.1). The plants were imaged from 6 different angles for the fluorescence, thermal infrared and RGB cameras: top down view, 0, 36, 72, 108 and 144 degrees.

A.3 Data description and potential use

In this experiment a total of 15 plants were grown. The plants represent different species, genotypes, geographical locations, NADP and NAD-ME C₄ photosynthetic subtypes and 4 independent origins of C₄ photosynthesis (Table A.1). The dataset is stored at CyVerse
under the doi 10.25739/7n5e-9w17. The folders for experiment A are organized in the following format: Species/genotype name -> Camera type (VIS, FLU, IR, HYP). The hyperspectral camera stored 243 images for each wavelength ranging from 546 nm to 1700 nm, with a pixel resolution of 320x561 between 11 October 2017 and 20 December 2017 and 320x498 between 21 December 2017 and 05 February 2018. The fluorescence camera stored images at wavelengths 500-700nm based on excitation, with light at 400-500nm and a pixel resolution of 1038x1390. The RGB camera collected images with pixel resolution of 2454x2056. For zoom level 1, there are 0.077 pixels per cm; for zoom level 2, there are 0.105 pixels per cm, and for zoom level 3, there are 0.175 pixels per cm.

Throughout the course of the four months, the zoom of the images captured by the RGB camera changed to make sure it matched the plant growth and stage. The zoom changes happened 24 November 2017, 15 December 2017 (Figure A.1).

The images collected here could potentially help the development of new crop growth models for proso, pearl, foxtail and Japanese millet and teff. Therefore these images can be used to match the predictions from the growth models for plant height and leaf count and area. The plant height obtained for all three maize genotypes were within the same range for the same time period found in a previous study (Figure A.2). Also, with the availability of crop models described for maize, these images could be analyzed using maize crop models to test how translatable is such model to the other grasses imaged here. This test allows us to assess how broad or specific the crop growth models for maize are, and if they could be used to predict the growth models of other related grasses in this study. The data obtained for both maize and sorghum can be used to test how well previous growth models of these plants predict what was collected in this study, helping validate such models. Obtaining image data for crops for a long period of time can be challenging, which could limit the amount of information available to test new methods. For this reason, the present dataset can be used in the development of new methods to estimate different plant traits.
This dataset comprises C₄ grasses belonging to the two of the three C₄ photosynthesis subtypes, NAD-ME and NADP-ME. C₄ photosynthesis can be categorized into main three distinct subtypes, depending on the decarboxylating enzyme is used to provide CO₂ to the Calvin cycle: NADP-ME, NAD-ME and PCK. The availability of image data from crops representing two different C₄ photosynthesis subtypes (Table A.1) is relevant to compare different photosynthetic plant traits and investigate whether the two subtypes have specific traits that could enhance plant growth or yield.

The folders for experiment B are organized in the following format: sorghum barcode name Camera type and angle (VIS, FLU, IR, HYP). For example: "Sorghum_Btx623R1WW-286//Fluo_SV_O". The bold part of the name indicates the date the plant was imaged. As in experiment A the RGB, fluorescence and thermal infrared cameras have an extra nesting level in the data structure, -> Angle (0, 36, 72, 108, 144, and top view). Images were collected for 30 days (Figure A.3). Images from experiment B can be used for sorghum crop growth models as in experiment A. In addition, the ground truth photosynthetic data collected for the sorghum plants could potentially allow the development of growth models that predict how this plant would grow under different environments. Besides image data, watering data is also available for this experiment. Therefore, it is possible to calculate how much water each plant imaged used and how much of it was lost due to evapotranspiration (Table SA.2).

**A.4 Limitations**

The LemnaTec system presented errors and did not image plants every other day, so there is missing data in both datasets. In the dataset generated for experiment A, plants were grown in the LemnaTec system for 115 days, where images were not taken in 11 days. Experiment B is missing 3 days of image data, January 18th, 24th and 28th. For the same experiment, there is missing data from 6 days of watering data on January 18th, 20th-23rd and 28th. There is missing data on 25 October 2017 (only sorghum), 22 November 2017 for
Figure A.1: RGB images collected for pearl millet, genotype C042, experiment A, between 13 October 2017 and 05 February 2018. Images in zoom 2 had their brightness adjusted manually.

experiment A. This level of missing data does not affect the possibility to use the images collected here to test crop growth models, or use in the development of new prediction methods.

Also, the zooming of hyperspectral data for taller plants (i.e. maize, sorghum and pearl millet) did not allow the retrieval the top part of the plants. Despite this limitation,
Figure A.2: Plant growth curves measured for all species in all angles through the period of 115 days after planting. Green dashed lines represent the zoom changes, showing the maximum height for that zoom level. Sudden drop in plant height of foxtail millet PI614815 was due the plant size, making it lean over. Plant was staked to avoid that problem.

Figure A.3: RGB images collected for sorghum BTx623, experiment B, between 01 January 2018 and 30 January 2018. Images from different angles were used as the position of pots were changed during the experiment.

the hyperspectral images still allow assessment of plant greenness and finding normalized digital vegetative index (NDVI) for detecting the plant pixels and area.
Another limitation of this data set is that we could not grow replicates for four months, which means only individual plants were imaged. Although this can be limiting for trait heritability studies, this limitation can be surpassed by using other datasets available for maize and sorghum. Also, the lack of replication does not impair heritability analysis between species rather than genotypes. For the experiment B sorghum plants were not imaged from seedling stage and could not be grown to maturity, which could be a limiting factor when developing statistical models for sorghum.

From January 20th to January 23rd, no pot weights or watering data was recorded for the sorghum plants in experiment B, however, images were still captured. Depending on the nature of the research being conducted, this missing data may not be a significant issue, or it may be necessary to impute the missing datapoints.209

### A.5 Supplemental material

<table>
<thead>
<tr>
<th>Pot position</th>
<th>Common name</th>
<th>Variety</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>Sorghum</td>
<td>BTx623</td>
</tr>
<tr>
<td>02</td>
<td>Pearl millet</td>
<td>C042</td>
</tr>
<tr>
<td>03</td>
<td>Japanese millet</td>
<td>BN-14698-63</td>
</tr>
<tr>
<td>04</td>
<td>Finger millet</td>
<td>I. E. 595</td>
</tr>
<tr>
<td>05</td>
<td>Teff</td>
<td>R286</td>
</tr>
<tr>
<td>06</td>
<td>Empty pot</td>
<td></td>
</tr>
<tr>
<td>07</td>
<td>Sorghum</td>
<td>CM052</td>
</tr>
<tr>
<td>08</td>
<td>Proso millet</td>
<td>Earlybird</td>
</tr>
<tr>
<td>09</td>
<td>Pearl millet</td>
<td>NM-6R1</td>
</tr>
<tr>
<td>10</td>
<td>Maize/corn</td>
<td>DE811</td>
</tr>
<tr>
<td>11</td>
<td>Maize/corn</td>
<td>Va99</td>
</tr>
<tr>
<td>12</td>
<td>Empty pot</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Foxtail millet</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Teff</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Foxtail millet</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>Maize</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>Sorghum</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>Empty pot</td>
<td></td>
</tr>
</tbody>
</table>

Table A.S1: Experimental layout.
<table>
<thead>
<tr>
<th>Day</th>
<th>Snapshot ID tag</th>
<th>Snapshot time stamp</th>
<th>Water amount</th>
<th>Weight after</th>
<th>Weight before</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1127-TX623Ae-WW-300 1127-TX623Ae-WW-305 1127-TX623Ae-WW-286</td>
<td>1/11/2018 15:26</td>
<td>2215</td>
<td>6874</td>
<td>4635</td>
</tr>
<tr>
<td>1</td>
<td>1127-TX623Ae-WW-300 1127-TX623Ae-WW-305 1127-TX623Ae-WW-286</td>
<td>1/11/2018 15:33</td>
<td>2428</td>
<td>6876</td>
<td>4422</td>
</tr>
<tr>
<td>1</td>
<td>1127-TX623Ae-WW-300 1127-TX623Ae-WW-305 1127-TX623Ae-WW-286</td>
<td>1/11/2018 15:03</td>
<td>2466</td>
<td>6887</td>
<td>4384</td>
</tr>
<tr>
<td>2</td>
<td>1127-TX623Ae-WW-300 1127-TX623Ae-WW-286</td>
<td>1/12/2018 3:49</td>
<td>76</td>
<td>6841</td>
<td>6774</td>
</tr>
<tr>
<td>2</td>
<td>1127-TX623Ae-WW-300 1127-TX623Ae-WW-286</td>
<td>1/12/2018 3:11</td>
<td>61</td>
<td>6848</td>
<td>6789</td>
</tr>
<tr>
<td>2</td>
<td>1127-TX623Ae-WW-300 1127-TX623Ae-WW-286</td>
<td>1/12/2018 4:01</td>
<td>68</td>
<td>6849</td>
<td>6782</td>
</tr>
<tr>
<td>3</td>
<td>1127-TX623Ae-WW-286</td>
<td>1/13/2018 3:11</td>
<td>83</td>
<td>6696</td>
<td>6617</td>
</tr>
<tr>
<td>3</td>
<td>1127-TX623Ae-WW-286</td>
<td>1/13/2018 3:49</td>
<td>102</td>
<td>6697</td>
<td>6598</td>
</tr>
<tr>
<td>3</td>
<td>1127-TX623Ae-WW-300 1127-TX623Ae-WW-305</td>
<td>1/13/2018 4:00</td>
<td>66</td>
<td>6699</td>
<td>6634</td>
</tr>
<tr>
<td>4</td>
<td>1127-TX623Ae-WW-305</td>
<td>1/14/2018 4:01</td>
<td>204</td>
<td>6690</td>
<td>6496</td>
</tr>
<tr>
<td>4</td>
<td>1127-TX623Ae-WW-286</td>
<td>1/14/2018 3:10</td>
<td>235</td>
<td>6705</td>
<td>6465</td>
</tr>
<tr>
<td>4</td>
<td>1127-TX623Ae-WW-300 1127-TX623Ae-WW-305</td>
<td>1/14/2018 3:49</td>
<td>242</td>
<td>6705</td>
<td>6458</td>
</tr>
<tr>
<td>5</td>
<td>1127-TX623Ae-WW-300 1127-TX623Ae-WW-286</td>
<td>1/15/2018 3:49</td>
<td>223</td>
<td>6689</td>
<td>6477</td>
</tr>
<tr>
<td>5</td>
<td>1127-TX623Ae-WW-286</td>
<td>1/15/2018 3:10</td>
<td>210</td>
<td>6693</td>
<td>6490</td>
</tr>
<tr>
<td></td>
<td>1127-TX623Ae-WW-305</td>
<td>1/15/2018 4:01</td>
<td>203</td>
<td>6706</td>
<td>6497</td>
</tr>
<tr>
<td>---</td>
<td>---------------------</td>
<td>----------------</td>
<td>-----</td>
<td>------</td>
<td>------</td>
</tr>
<tr>
<td>5</td>
<td>1127-TX623Ae-WW-300</td>
<td>1/16/2018 3:49</td>
<td>0</td>
<td>6405</td>
<td>6405</td>
</tr>
<tr>
<td>6</td>
<td>1127-TX623Ae-WW-286</td>
<td>1/16/2018 3:10</td>
<td>0</td>
<td>6418</td>
<td>6418</td>
</tr>
<tr>
<td>6</td>
<td>1127-TX623Ae-WW-305</td>
<td>1/16/2018 4:00</td>
<td>0</td>
<td>6482</td>
<td>6482</td>
</tr>
<tr>
<td>6</td>
<td>1127-TX623Ae-WW-300</td>
<td>1/17/2018 6:49</td>
<td>0</td>
<td>6108</td>
<td>6109</td>
</tr>
<tr>
<td>7</td>
<td>1127-TX623Ae-WW-286</td>
<td>1/17/2018 6:10</td>
<td>0</td>
<td>6134</td>
<td>6134</td>
</tr>
<tr>
<td>7</td>
<td>1127-TX623Ae-WW-305</td>
<td>1/17/2018 7:00</td>
<td>0</td>
<td>6243</td>
<td>6243</td>
</tr>
<tr>
<td>7</td>
<td>1127-TX623Ae-WW-286</td>
<td>1/19/2018 3:13</td>
<td>460</td>
<td>6091</td>
<td>5640</td>
</tr>
<tr>
<td>9</td>
<td>1127-TX623Ae-WW-300</td>
<td>1/19/2018 3:51</td>
<td>567</td>
<td>6095</td>
<td>5533</td>
</tr>
<tr>
<td>9</td>
<td>1127-TX623Ae-WW-305</td>
<td>1/19/2018 4:03</td>
<td>307</td>
<td>6104</td>
<td>5793</td>
</tr>
<tr>
<td>10</td>
<td>1127-TX623Ae-WW-286</td>
<td>1/20/2018 3:12</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>10</td>
<td>1127-TX623Ae-WW-300</td>
<td>1/20/2018 3:50</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>10</td>
<td>1127-TX623Ae-WW-305</td>
<td>1/20/2018 4:02</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>11</td>
<td>1127-TX623Ae-WW-286</td>
<td>1/21/2018 3:12</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>11</td>
<td>1127-TX623Ae-WW-300</td>
<td>1/21/2018 3:50</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>Date</td>
<td>Time</td>
<td>Disk</td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>------------</td>
<td>--------</td>
<td>------</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>11</td>
<td>1/21/2018</td>
<td>4:02</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>12</td>
<td>1/22/2018</td>
<td>3:12</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>12</td>
<td>1/22/2018</td>
<td>3:49</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>12</td>
<td>1/22/2018</td>
<td>4:01</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>12</td>
<td>1/23/2018</td>
<td>4:11</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>12</td>
<td>1/23/2018</td>
<td>4:49</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>12</td>
<td>1/23/2018</td>
<td>5:00</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>15</td>
<td>1/25/2018</td>
<td>17:09</td>
<td>1483</td>
<td>5790</td>
<td>4317</td>
</tr>
<tr>
<td>15</td>
<td>1/25/2018</td>
<td>17:13</td>
<td>1178</td>
<td>5793</td>
<td>4622</td>
</tr>
<tr>
<td>15</td>
<td>1/25/2018</td>
<td>17:00</td>
<td>1239</td>
<td>5797</td>
<td>4561</td>
</tr>
<tr>
<td>16</td>
<td>1/26/2018</td>
<td>2:49</td>
<td>102</td>
<td>5799</td>
<td>5698</td>
</tr>
<tr>
<td>16</td>
<td>1/26/2018</td>
<td>2:10</td>
<td>78</td>
<td>5803</td>
<td>5722</td>
</tr>
<tr>
<td>16</td>
<td>1/26/2018</td>
<td>3:00</td>
<td>89</td>
<td>5806</td>
<td>5711</td>
</tr>
<tr>
<td>17</td>
<td>1/27/2018</td>
<td>2:11</td>
<td>289</td>
<td>5789</td>
<td>5511</td>
</tr>
<tr>
<td>17</td>
<td>1/27/2018</td>
<td>3:00</td>
<td>276</td>
<td>5789</td>
<td>5524</td>
</tr>
<tr>
<td></td>
<td>1127-TX623Ae-WW-300</td>
<td>1/27/2018 2:49</td>
<td>339</td>
<td>5795</td>
<td>5461</td>
</tr>
<tr>
<td>----</td>
<td>------------------</td>
<td>----------------</td>
<td>------</td>
<td>------</td>
<td>------</td>
</tr>
<tr>
<td>19</td>
<td>1127-TX623Ae-WW-300</td>
<td>1/29/2018 15:14</td>
<td>920</td>
<td>5782</td>
<td>4880</td>
</tr>
<tr>
<td>19</td>
<td>1127-TX623Ae-WW-305</td>
<td>1/29/2018 15:17</td>
<td>773</td>
<td>5786</td>
<td>5027</td>
</tr>
<tr>
<td>19</td>
<td>1127-TX623Ae-WW-286</td>
<td>1/29/2018 15:05</td>
<td>837</td>
<td>5791</td>
<td>4963</td>
</tr>
<tr>
<td>20</td>
<td>1127-TX623Ae-WW-305</td>
<td>1/30/2018 3:00</td>
<td>134</td>
<td>5797</td>
<td>5666</td>
</tr>
<tr>
<td>20</td>
<td>1127-TX623Ae-WW-286</td>
<td>1/30/2018 2:10</td>
<td>131</td>
<td>5805</td>
<td>5669</td>
</tr>
<tr>
<td>20</td>
<td>1127-TX623Ae-WW-300</td>
<td>1/30/2018 2:49</td>
<td>162</td>
<td>5805</td>
<td>5638</td>
</tr>
</tbody>
</table>

Table A.S2: Summary of watering data of experiment B. Values "-1" were replaced for "NA" values, which represent missing data.