C4 PHOTOSYNTHETIC EVOLUTION:

SUB-TYPES, DIVERSITY, AND FUNCTION WITHIN THE GRASS TRIBE PANICEAE

A Dissertation

presented to

the Faculty of the Graduate School

at the University of Missouri-Columbia

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

---

by

JACOB DANIEL WASHBURN

Dr. J. Chris Pires, Dissertation Supervisor

MAY 2017

The undersigned, appointed by the dean of the Graduate School, have examined the dissertation entitled:

C4 PHOTOSYNTHETIC EVOLUTION: SUB-TYPES, DIVERSITY, AND FUNCTION WITHIN THE GRASS TRIBE PANICEAE

Presented by Jacob Daniel Washburn, a candidate for the degree of doctor of philosophy, and hereby certify that, in their opinion, it is worthy of acceptance.

_____
Dr. J. Chris Pires

_____
Dr. James A. Birchler

_____
Dr. Paula McSteen

_____
Dr. Gavin Conant

ACKNOWLEDGEMENTS

I would first like to thank my beautiful wife and sweetheart Melinda for her constant companionship, support, and sacrifice over the past five years. Also my three children: Nathan, Sam, and Emma. The four of you have been, and continue to be my inspiration, and my happiness. I also want to thank my parents, Shelley and Kevin Washburn, who instilled in me a love for learning and for hard work. This degree is for you as well.

I also thank my advisor Chris for being the most supportive, helpful, and forward-thinking mentor I have ever had the privilege of associating with. I credit you with the success I have had in grant writing during my Ph.D., and with many of the life skills I have learned. My co-advisor Jim has also been an incredible help. Jim, in my mind you are the epitome of a scientist. I also want to thank Gavin for serving on my committee, and for mentoring me in writing, and in thinking like a scientist. Paula, thank you for mentoring me while we served together on the hiring committee, as well as being a crucial member of my dissertation committee. I also wish to thank the Division of Biological Sciences' staff, students, and faculty for providing such a stimulating and supportive environment.

Lastly, I want to thank my church community for their support, and my Heavenly Father for enabling and assisting me in accomplishing this great task.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

ABSTRACT

Most plants convert sunlight into chemical energy using a process known as C_3 photosynthesis. However, some of the world's most successful plants instead use the C_4 photosynthetic pathway which allows them to more efficiently use water, nitrogen, and solar energy. In the past 30 million years, C_4 photosynthesis has convergently evolved from C_3 over 60 times and new lineages are in the process of evolving even today. Because of this complex evolutionary history, C_4 is not "one" uniform photosynthetic type, but a diverse collection of photosynthetic sub-types that are classically grouped according to their use of three different biochemical pathways. The grass tribe Paniceae is especially interesting in this aspect because it contains all three of these biochemical sub-types as well as important food and bioenergy crops.

To better understand the evolution of C_4 photosynthesis, DNA and RNA sequencing were undertaken for various species from within the Paniceae and used for phylogenetic and comparative genomic studies. Cell type specific RNA expression profiling for the two major C_4 cell types was also completed for representative species of each C_4 sub-type. Streamlined bioinformatics pipelines for both chloroplast and nuclear phylogenetics were developed for processing the data. These analyses resulted in: 1) The first "genome scale" phylogenetic tree of the grass tribe Paniceae, 2) The clearest evidence to date of the evolutionary relationships between the three classically defined C_4 sub-types, 3) The most convincing results to date that the chloroplast and nuclear phylogenies of the Paniceae are incongruent, 4) Evidence that this chloroplast nuclear incongruence is likely due to introgression and/or incomplete lineage sorting, and 5) Strong support for sub-type mixing as well as the existence of a PCK sub-type.

CHAPTER 1: CONVERGENT EVOLUTION AND THE ORIGIN OF COMPLEX PHENOTYPES IN THE AGE OF SYSTEMS BIOLOGY

Jacob D. Washburn[1], Kevin A. Bird[1], Gavin C. Conant[2] and J. Chris Pires[1,3]

1 Division of Biological Sciences, University of Missouri, 311 Bond Life Sciences Center, Columbia, Mo 65211 USA
2 Division of Animal Sciences & Informatics Institute, University of Missouri, Columbia, MO 65211 USA
3 Informatics Institute & Bond Life Sciences Center, University of Missouri, 371b Bond Life Sciences Center, Columbia, Mo 65211 USA

Please cite the published version in the International Journal of Plant Sciences:

Washburn JD, KA Bird, GC Conant, JC Pires 2016 Convergent evolution and the origin of complex phenotypes in the age of systems biology. Int J Plant Sci 177: 305-318. http://www.journals.uchicago.edu/doi/abs/10.1086/686009

Abstract

Convergent evolution has fascinated and occasionally mystified biologists since the principle of universal common ancestry was accepted. Similar phenotypes can arise by common ancestry (including pre-adaptations) or through constraints in the space of possible phenotypes, and can increase in a population either via drift or selection. Assessing which of these mechanisms to invoke for any given example remains challenging for both simple and complex phenotypes. However, barriers in this area are slowly breaking down with recent advances in genomics and systems biology. A renaissance in the study of convergent evolution may be on its way, as surprising explanations for similar phenotypes, such as the metabolic similarities between yeast and cancer cells, are uncovered with network and metabolic models. We argue that although examples of convergence are known from many domains of life, green plants in particular have remarkable promise for the study of convergence because they are experimentally tractable, have considerable –omics and systems biology resources available, and show convergence in a number of important and complex traits. Four such examples include the "domestication syndrome", duplicate loss and retention patterns following whole genome duplication, the multiple appearances of C_4 and CAM photosynthesis, and hybrid vigor.

Introduction

Convergent evolution is the appearance of similar phenotypes in distinct evolutionary lineages. Common examples include the independent origins of flight in insects, birds, and mammals and the multiple appearances of the camera-like eye in lineages such as cephalopods and vertebrates (Futuyma, 1998; Gehring and Ikeo, 1999). Some lesser-known, but perhaps even more intriguing, examples of the convergent evolution of complex traits include patterns of gene retention and loss after whole genome duplication (WGD), the recurrent appearances of C_4 and CAM photosynthesis in plants, the "domestication syndrome" in both plants and animals, and cross-kingdom examples of hybrid vigor.

Biologists and philosophers have proposed a variety of definitions of convergent evolution, with considerable discussion of the precise meaning of terms such as convergent evolution and parallel evolution (Pearce, 2012; Currie, 2013). While recognizing the importance of these definitions, we here offer a broad look at evolutionary events involving the independent origins of complex traits. We will use *convergence* and *convergent evolution* as umbrella terms to encompass phenomena that might also be referred to as parallel evolution. Thus, when we use the term convergent evolution, we are not making the claim that the homoplasious phenotype in question was shaped by selection in the lineages involved, but merely that similar phenotypes exist in those lineages in a manner that cannot be easily explained by descent from a common ancestor.

Theories Explaining Convergent Evolution of Similar Phenotypes

An illuminating, though imperfect, metaphor for the study of convergent evolution is Gould's thought experiment of "replaying the tape of life" (Gould, 1989), an idea, incidentally, with antecedents going back at least to Fisher (1934). Gould's question is: if the tape of life were replayed, would the results (namely the different modern biological forms on the planet) be the same or different from those we see today? Gould himself argued that, due to the contingent nature of evolution, each replay of the tape would result in a different outcome. Others, such as Simon Conway Morris, have argued that evolution is strongly shaped by constraints: physical or developmental limitations on the number of solutions to a given problem (Conway Morris, 2003). These constraints force evolution to take specific, recurring paths, meaning that replaying the tape would result in *similar*, but not identical, outcomes each time.

Cases of convergent evolution are sometimes thought of as examples of the "tape of life" being replayed with similar results. While this analogy is useful and some convergent events may have in fact been influenced by similarities in evolutionary history, these events by definition do not share the exact same history. More importantly, the mere appearance of similarity does not prove that selection, operating under constraints, has driven the common outcome.

We propose that at least three evolutionary processes (aside from common ancestry itself) can give rise to similar phenotypes: 1) similar selective forces may drive a trait's development in multiple lineages; 2) underlying constraints may make the trait's evolution inevitable or highly probable under certain conditions; and/or 3) the trait's repeated emergence may be attributable to genetic drift (See Figure 1.1). These three

processes are of course not mutually exclusive: indeed most cases of convergence may result from a mixture of all three. Hence, it can be informative to consider these processes individually and to examine their relative importance in different examples of convergence.

*Selection***.** Convergence by selection occurs when similar selective forces acting on distinct lineages result in similar traits within those lineages. C_4 photosynthesis in plants (discussed in more detail later) is one example of selection's role in convergent evolution. In this case, different enzymes, biochemical pathways, and anatomical configurations are employed by different lineages of C_4 plants; nonetheless, those lineages seem to have evolved under similar selective pressures (Sage et al., 2012). Another compelling example of convergence by selection is the crystalline lenses of birds and mammals. These lenses perform similar functions but are made from different proteins and minerals, and are constructed by different enzymes in different lineages (Schwab, 2012; Map of Life, 2015). Other examples indicating a strong role for selection can be found across kingdoms and include the independent recent evolution of nylonases (Prijambada et al., 1995) and the evolution of novel glycolytic enzymes from enzymes with other functions in an *E.coli* glycolytic knock-out strain (Miller and Raines, 2004, 2005). In each of these examples there is a clear role for selection, be it increased photosynthetic efficiency, better sensory perception, or the ability to survive on a particular carbon source. In each case there is also clear evidence that more than one viable solution to the problem exists, since the outcomes are similar at the gross level but differ in their details.

Constraint. Physical, biochemical, or developmental constraints that reduce the pool of potential genetic solutions to a given problem are additional mechanisms that can yield similar phenotypes along independent evolutionary trajectories. The word constraint is often used in the context of selective constraint, meaning that a specific trait may arise through mutation but be quickly removed from the population by selection. Here we use the terms differently, namely to refer to a trait that simply cannot come into existence because of limitations due to physics, chemistry, and/or evolutionary history. In its simplest form this viewpoint suggests that, in some cases, there are a limited number of possible ways to achieve a given outcome (Weinreich et al., 2006). For example, there are no six-legged mammals. This deficit is most likely not because six-legged mammals have at some point evolved and been selected against, but simply because the ancestors of mammals were tetrapods and the genetic architecture necessary to create a six-legged mammal does not exist.

Other examples of convergent evolution by constraint include similar metabolic traits arising under experimentally controlled circumstances (Ibarra et al., 2002), and similarities between human-designed networks like the U.S. power grid and biological networks (Watts and Strogatz, 1998; Milo et al., 2002). These similarities can take the form of networks structured in such a way that they show similar statistics like centrality of the networks or the recurrence of common motifs whereby the same connection pattern among a small group of nodes occurs many times. These similar structures also tend to make the networks all show "small world" features, meaning that any node in the network can be reached by traversing a small number of other nodes (6 or less) (Watts and Strogatz, 1998; Wagner and Fell, 2001). These studies suggest that there are a

limited number of ways to successfully build a network, be it biological or otherwise, with desirable properties (Jeong et al., 2001).

Drift. Genetic drift can also explain many examples of convergent evolution. The antifreeze proteins of arctic and Antarctic fishes, for example, probably started with such mutations that were later co-opted and expanded (Chen et al., 1997; Fletcher et al., 2001). Since examples of convergence that are explained by drift alone are perhaps less interesting and/or pertinent to the focus of this review, we note their existence and importance but will leave their discussion to other articles.

As noted earlier, these three mechanisms (selection, constraint, and drift) for achieving similar phenotypes are not mutually exclusive, and roles for each can be found in many examples of convergence. More importantly, it is not always easy or even possible to tease apart convergent evolution from common ancestry. For instance, pre-adaptations in a lineage (initially non-adaptive traits with the potential for co-option as adaptions later in their history) may make that lineage more likely to evolve a given phenotype. This recurrence is not due to the particular architecture of the phenotype being the optimal solution to the problem but to the fact that the building blocks for that trait architecture are already in place. Symbiosis between plants and nitrogen fixing bacteria through nodulation is one such example. It occurs intermittently across multiple families within angiosperms. While diversity exists between many of these apparently isolated occurrences, there have also been suggestions that they all come from a common ancestor with a predisposition to symbiosis (Soltis et al., 1995; Huss-Danell, 1997; Doyle and Luckow, 2003; Delaux et al., 2015; Li et al., 2015b). An additional example is the compound eye, which has independently evolved several times (Futuyma, 1998; Gehring

and Ikeo, 1999), but whose independent evolutionary origins all build on photosensitivity, an ancient animal trait with a common set of sensory and developmental genes shared by most metazoans. The power of these shared developmental programs is illustrated by the fact that the mouse homolog of the *Drosophila eyeless* gene is able to induce ectopic eye development in *Drosophila* (Quiring et al., 1994; Halder et al., 1995). Hence, while eye morphology has evolved convergently (multiple origins of camera-like eyes and compound eyes), the deep origin of eyes is built, at least in part, from of a common inherited set of genes for light perception. C_4 photosynthesis in plants is another example where pre-adaptation may have played a significant role (see C_4 photosynthesis section below). These examples serve as an important cautionary reminder of the role common ancestry can play in a trait's emergence, even when that role may not be readily apparent.

Challenges in Differentiating Among Causes of Convergence

Some basic guidelines can be helpful in determining the causes of convergent evolution, particularly when the genetic underpinnings of a trait are sufficiently well known. From a genetic perspective, adaptive convergences may have arisen through two basic scenarios. In the first, a mutation or mutations in the same gene or genes caused the homoplasy in the organisms. In the second the causal mutation or mutations occurred in different genes in each lineage (Wake et al., 2011). Determining which scenario occurred for a particular trait can suggest the degree to which each of the three above mechanisms of convergence is operating. If the same gene(s) have given rise to similar phenotypes in

independent lineages ("gene reuse"), this suggests that constraint or pre-adaptation may have played an important role in the convergent phenotype (Conte et al., 2012). In contrast, evidence for selection without constraint comes from different genes contributing to the same trait in different lineages (Losos, 2011).

Examples of gene reuse are strikingly common in domestication, implying a small suite of genes might be responsible for domestication-related phenotypes in many organisms (Paterson et al., 1995; Ramsay et al., 2011; Butelli et al., 2012; Lenser and Theißen, 2013; Martin and Orgogozo, 2013). However, strong selection is also likely to have been important in domestication (at least in annual plants and most animals): in fact, some studies have not been able to detect constraints on gene reuse, arguing more strongly for selection as the primary factor (Gaut, 2015).

In this same vein, several examples of convergence in protein function also suggest a role for constraints. For instance, the co-option of lysozymes for digestion in foregut-fermenting herbivores seems to have evolved multiple times in a wide range of species, including monkeys, birds, insects, and bovids (Stewart et al., 1987; Kornegay et al., 1994; Regel et al., 1998). This type of convergence in protein function suggests that some amount of constraint is involved: however, caution is in order because evolution also tends to use the materials "closest to hand," meaning that lysozymes may simply have been particularly convenient for reuse in digestion.

Occasionally, proteins may converge not only in function but also in specific regions of their sequences. There are a number of examples of sequence-level convergence in both DNA and protein sequences (Zhang and Kumar, 1997; Soltis and Soltis, 1998; Soltis et al., 1999; Kriener et al., 2000; Li et al., 2008; Castoe et al., 2009;

Liu et al., 2010; Tenaillon et al., 2012; Parker et al., 2013; Stern, 2013; Natarajan et al., 2015). Castoe et al. (2009), found significant convergent molecular evolution in the amino acid substitutions observed between the mitochondrial genomes of snakes and agamid lizards, and Kriener et al. (2000), found convergence in the use of shared peptide motifs in the major histocompatibility complexes of humans and new world monkeys. In another example involving both plants and animals, Maier et al. (2013) demonstrate that the genes encoding ribosomal proteins are convergent in mitochondria and chloroplasts across all eukaryotes.

Two recent studies of convergent evolution in mammalian genomes serve to illustrate the challenges inherent in differentiating among the causes of convergence. The first study examined the convergent evolution of echolocation in bats and dolphins looking for evidence that different lineages not only share a convergent phenotype but also have experienced convergence at the sequence level (Parker et al., 2013). These authors found evidence of convergent sequence evolution in several hundred genes linked to either echolocation or vision. They were also able to demonstrate that many of these sequence changes showed signatures of selection (Parker et al., 2013). Foote et al. (2015) found similar results in an analysis of the convergent evolution of adaptations to marine environments across different mammalian orders. Their analyses paint the same picture as those of Parker et al. in that many genes appear to have evolved convergently at an amino acid sequence level and also appear to have experienced positive selection. However, unlike the previous study, these authors also applied their methods to a set of control species lacking adaptations to marine environments. Surprisingly, they found

many of the same patterns in the non-adapted species, suggesting that these tests of convergence at the sequence level may be somewhat prone to false positives.

Convergence and Complex Traits

The above collection of examples and possible sources of convergence raise the question of the best systems for understanding convergent evolution. While a variety of relatively simple traits have arisen multiple times independently, these can often be explained easily by common underlying genetics, genetic drift, or other simple evolutionary processes. For example, instances of specific gene regulatory circuit topologies such as feed forward loops have evolved multiple times independently within both yeast and *E.coli* (Conant and Wagner, 2003), but this recurrence is perhaps not surprising given their relatively simple structure. The independent origins of enzyme activities on the other hand are less susceptible to mutational explanations, but the common observation of enzyme promiscuity, the ability of an enzyme with one primary function to perform another function at some low level (O'Brien and Herschlag, 1999; Copley, 2003) makes the degree of selection at work somewhat obscure. These kinds of details can make the use of simple traits less meaningful to the study of convergence.

Because of these challenges with simple traits, it is probably more fruitful to study convergence in traits which appear to be complex in nature and are difficult to explain by common ancestry or simple evolutionary processes (Currie, 2013; Zaman et al., 2014). It would appear that unlike simple traits, the convergent evolution of relatively complex traits likely requires a combination of selection for a certain adaptation and a limit in the

space of forms available to be selected upon. In addition, the fact that many of the most important economic, agricultural, and medical phenotypes can be considered complex makes the study of convergent evolution of complex traits of significant general interest.

That said, it can be difficult to delineate a complex trait from a simple one, and how complex something is often depends on the scale at which one is looking. As we gain greater understanding of a given complex trait, we may find that it is actually very simple and that the surprising convergence we originally saw is really not so surprising. The examples of the compound eye and C_4 photosynthesis noted above are both traits with considerable complexity from a birds-eye view but for which the actual complexity of the innovation in question may be less dramatic when fully understood.

Over the years, different approaches have been developed and used for studying convergent evolution, particularly for complex traits. In the next section, we review the current tools for studying convergence before turning to future directions, novel tools, and interesting examples of complex convergence. For many of these examples, their in-depth dissection is only now becoming possible with recent theoretical and methodological advances. Because complex traits are by definition complex, we advocate for the increased use of systems biology tools (here defined broadly as any analysis involving multiple players and their interactions) in their study.

Approaches for Analyzing Convergent Evolution – Past and Present

The first step in deciphering if patterns of trait evolution are due to common ancestry or convergent evolution is often to put traits in the context of a phylogenetic tree. Once the phylogenetic pattern of convergence has been shown, the drivers of

convergence can be explored using the tools of evolutionary development (evo-devo) and comparative genomics. Other experimental approaches are also useful for demonstrating convergent evolution in ways which, to some extent, "replay the evolutionary tape" in a controlled setting.

Phylogenetics

In general, the study of convergence starts with a phylogenetic tree because a well supported and resolved phylogeny is necessary for the identification of traits that may be convergent versus those that arose by common decent. Hypotheses about which traits are convergent can be formed and some degree of support for those inferences can be made directly from the phylogeny. The strengths of phylogenetic tools for inferring convergence also include their utility in identifying probable transition points between traits, their amenability to statistical testing, and their ability to clearly support or refute hypotheses of convergence. For example, phylogenetic analyses have been instrumental in identifying the many distinct origins of C_4 and CAM photosynthesis (see discussion below).

Phylogenetics does have its limitations however. For example, it is generally unable to address the causes of the convergent events (i.e., selection, constraint, drift). Further tools that can describe such causes (e.g., comparative genomics, evolutionary developmental genetics, detailed morphological studies, experimental studies and ecological studies) often explicitly use phylogenies and their underlying statistical models as a starting point (Nielsen, 2005).

Evolution of Development (Evo-devo)

One important group of tools for understanding convergence is referred to as the study of the evolution of development or evolutionary development (evo-devo). These methods use comparisons between organisms and across developmental stages to infer the ancestral states that led to the developmental differences currently seen between organisms. Some of the most important contributions of evo-devo methods have to do with their ability to identify if a homoplasious trait is truly convergent. For example, one can determine if a trait is truly novel within a lineage by finding out how similar that trait is to another trait at various points along both of their developmental trajectories (Glover et al., 2015). Two structures that look very different from one another at maturity may actually arise from common morphology and/or molecular underpinnings at an earlier stage. To determine if this is the case, one can use detailed developmental analysis across multiple phylogenetic origins of a trait to determine if the trait has arisen multiple times from the same ancestral state or from different ancestral states. Over the years, evo-devo approaches have contributed significantly to a basic understanding of the convergent evolution of floral morphology in angiosperms (Preston and Hileman, 2009; Christin et al., 2010b; Glover et al., 2015). Floral morphology is incredibly variable across angiosperms, yet most descriptive work is done on mature flowers. In this context, what might looks like a petal in one species could be referred to as a completely different structure in another. To accurately understand how these two traits evolved, one needs to know that they share a common ancestry that is not apparent at maturity. By looking at the two traits at earlier developmental time points one can find the precursor form(s) from which both originate. These types of discoveries are aided by the use of labeling

techniques that allow one to find out if the same proteins exist in both precursory organs and if they are present at similar levels. Similarly, two organs that arose from different precursors could be identified as convergent because they both express the same genes at similar levels. One example of this kind of convergence is found in multiple origins of bilateral symmetry via parallel recruitment of TCP transcription factors across core eudicots (Preston and Hileman, 2009).

Comparative Genomics

Expansive collections of genomic data now allow for the analysis of the genomic variations responsible for morphological traits of interest (Tenaillon et al., 2012; Martin and Orgogozo, 2013). As the number of sequenced genomes rises, researchers are better able to carry out comparative analysis over larger phylogenetic distances and dissect the molecular basis of convergence. Improvements in the quality of sequenced genomes and their annotations also play a critical role in what kinds of analyses can be performed. Comparative genomic methods have proven useful in finding the molecular causes of convergent traits in a variety of plants, insects, and animals (Paterson et al., 1995; Parker et al., 2013; Denoeud et al., 2014). For example, quantitative trait loci (QTL) analysis allowed Paterson et al. (1995) to identify convergent domestication-related traits between sorghum, rice and maize. Denoud et al. (2014) found convergence in caffeine synthesis within eudicots. Likewise, as already discussed, Parker et. al. (2013) identified over 200 loci involved in the evolution of echolocation in both bats and dolphins on the basis of

the analysis of 22 genome sequences. Further studies of this scale are becoming possible in an increasing number of plant lineages.

Beyond identifying the genetic basis of convergent traits, comparative genomics can also aid in differentiating between the causes of convergence as described earlier. Large-scale genomic data can allow one to look for signatures of selection across multiple organisms and multiple genes with relative ease. These signatures, along with the actual genomic sequences underlying the traits allow one to make strong inferences about the roles of constraint and selection in the convergence of a trait.

In addition to parsing the contributions of selection and constraint in convergent traits, comparative genomics allows for the study of convergent evolution in the genomes themselves. For example, convergent genome shrinkage, the loss of large amounts of similar types of genomic DNA, has been linked to whole-genome duplication (WGD) (discussed in more detail in a later section) and endosymbiosis (Paterson et al., 2006; van Hoek and Hogeweg, 2007; McCutcheon et al., 2009; McCutcheon and Moran, 2010).

Experimental Approaches

Various experimental approaches have also been used to elucidate the forces driving convergent phenotypes. The main class of experimental approaches that have been employed involve "replaying the tape of life" in the laboratory (Kawecki et al., 2012; Lobkovsky and Koonin, 2012; Barrick and Lenski, 2013; Matos et al., 2015). These approaches are of course limited to organisms that are amenable to a lab environment and experiments that can be performed in a reasonable amount of time. The

example of re-evolving a glycolysis enzyme in *E. coli* using substrate ambiguity has already been mentioned (Miller and Raines, 2004, 2005). By applying artificial selection, the researchers were able to select for a spontaneous mutation that increased a gene's expression to the point that it could restore function to a knocked-out pathway (Miller and Raines, 2005). Another example in *E. coli* is the evolution of tolerance to heat stress, where constraints appear to play an important role in the convergence of genes (Gaut, 2015). Similar experimental protocols have been employed in fruit flies; convergence was found in the frequency of alleles in populations under the same selection regime (Burke et al., 2010; Kawecki et al., 2012) as well as in phenotypic traits when distinct populations are placed under similar selection (Fragata et al., 2014). Other examples of experimental evolution have studied bacteriophages and shown repeatability in the way the virus evolves to attack its host (Meyer et al., 2012).

Other types of experimental approaches have also been used to elucidate the forces driving convergent phenotypes, including functional analyses of behavior. One example of this comes from studies of undulation swimming, which has independently evolved to a mechanical optimum in multiple phylogenetically-distant species from flatworms to rays to knifefish (Bale et al., 2015). In this example it appears that an optimum swimming pattern exists and any large departure from that pattern is strongly selected against (Bale et al., 2015). If there are any local optima, they are small enough and selection pressure is strong enough that they appear to have been quickly outcompeted by the global optimum. The researchers demonstrated this fact experimentally by first phylogenetically identifying homoplasy in the undulation swimming phenotype. Then, they employed computational and experimental modeling

to explore how mechanical swimming efficiency changes under perturbations to the observed natural patterns. These experiments were able to elegantly demonstrate that a mechanical optimum does in fact exist for this trait and variation away from that optimum drastically reduces efficiency. These results make a strong case for an interaction between strong selective forces and a narrow (constrained) pool of potential solutions to the problem.

One last example of using experimental evolution to study convergence in plants comes from the re-synthesis of 50 *Brassica napus* allopolyploids (Gaeta et al., 2007; Gaeta and Pires, 2010; Xiong et al., 2011). These lines were maintained for over ten generations and analyzed for structural and expression level changes over that time. Although the evolution of the lines post-allopolyploidization involved many gene losses and even the loss and replacement of whole chromosomes, these losses were shown to have more recurrent events than explicable by chance (Gaeta et al., 2007; Xiong et al., 2011). The patterns of homoeologous recombination found in these resynthesized lines were also found in natural accessions of *Brassica napus* (Chalhoub et al., 2014) .

A case study in convergence: C_4 photosynthesis

Plant biologists are also integrating an array of tools to study convergent evolution, as exemplified by studies of C_4 photosynthesis. Most land plants use the ancestral form of photosynthesis, known as C_3, to harness energy from sunlight for the conversion of carbon dioxide into sugars and other products (Sage et al., 1999; Sage et al., 2012). Around three percent of plants have evolved additional mechanisms that

concentrate carbon dioxide around the enzyme RuBisCO (Ribulose-1,5-bisphosphate carboxylase/oxygenase), thereby increasing the efficiency of photosynthesis (Sage et al., 1999; Sage et al., 2012). These mechanisms are known broadly as C_4 photosynthesis.

C_4 photosynthesis is generally considered a complex trait because many changes are required to go from C_3 to C_4 (Sage et al., 2012). For example, C_4 uses a modified anatomical leaf structure, often called Kranz anatomy, enzymatic pathways not generally found in C_3 plants, and a suite of gene expression and metabolic changes on both quantitative and qualitative levels. Even so, there are suggestions that certain lineages may be pre-adapted for C_4 evolution, and evidence is accumulating that the anatomical changes required for C_4 may be simpler than once thought (Grass Phylogeny Working Group II, 2012; Slewinski et al., 2012; Slewinski, 2013; Cui et al., 2014).

On the basis of the angiosperm phylogeny, the C_4 phenotype has independently evolved over 60 times (Sage et al., 2012). Within grasses (family Poaceae) alone, the trait seems to have evolved at least 22 times (Grass Phylogeny Working Group II, 2012) (See Figure 1.2). In some of these cases, well-supported phylogenies with good species sampling show C_4 clades that are well separated from each other by C_3 clades and long periods of evolutionary time. In these cases, a parsimonious reconstruction of the trait's evolution on the phylogeny clearly favors multiple C_4 origins over a single C_4 origin and multiple reversions to C_3. However, when one examines more closely related species that show a mix of C_3 and C_4 photosynthetic systems, it becomes less clear whether these apparently novel C_4 origins are indeed convergent appearances of C_4 or if instead they are only apparent reappearances that are in fact due to reversions back to C_3.

A number of authors have sought to resolve this difficulty using grasses as a test case. On the basis of both phylogenetic reconstructions of ancestral states and biochemical, anatomical, and genetic studies (Christin and Besnard, 2009; Christin et al., 2010a; Roalson, 2011; Grass Phylogeny Working Group II, 2012), these researchers have shown that the rate of C_3 to C_4 conversions within grasses is probably many times higher than the rate of conversions from C_4 back to C_3 (Grass Phylogeny Working Group II, 2012). One source of evidence for this bias is that the remnants of the C_4 pathway do not appear to be retained within C_3 species as would be expected in the case of a reversion from C_4 back to C_3. If this observation holds more generally (outside of grasses), then nearly all examples of C_4 lineages separated by C_3 lineages are in fact likely to be convergent and not merely due to a high rate of reversion back to the C_3 phenotype.

The case for convergence of C_4 photosynthesis is bolstered by a partial understanding of the selective regimes likely to favor that phenotype. Researchers believe that most C_4 origins result from selection for the prevention of photorespiration, a highly inefficient process which takes place in plants when CO_2 levels are low, temperatures are high, and/or water supplies are limited (Sage et al., 2012). Nonetheless, even with this extensive history of research on C_4 photosynthesis, our understanding of why it evolved so many times and the underlying causes are still quite imperfect. Open areas for investigation include: whether certain lineages were pre-adapted for C_4 photosynthesis, the reasons behind the evolution of different C_4 subtypes (Washburn et al., 2015), and the biological level (gene, cellular compartment, tissue) at which selection has acted for convergent C_4 structures. Because of the complex nature of C_4, systems biology tools should play a critical role in understanding its evolution. Several studies on C_4 to date

have applied tools that could be thought of as acting at a systems level, and they have made important discoveries about C_4 efficiency and the evolutionary path from C_3 to C_4 photosynthesis (de Oliveira Dal'Molin et al., 2010; Bräutigam et al., 2014; Heckmann et al.; Mangan and Brenner, 2014; Wang et al., 2014). Systems approaches have also been applied to understanding the evolution and development of C_4 leaf anatomy (de Oliveira Dal'Molin et al., 2010; Fouracre et al., 2014). However, these studies have likely only just begun to illustrate what systems biology can teach us about C_4.

Future Prospects for Investigating Convergent Evolution

What does the future hold for the study of convergent evolution, particularly in plants? What new tools are available now or will soon be available which may shed light on this phenomenon? What can plant systems tell us about convergence that others may not be able to?

Convergent Evolution as Seen Through the Lens of Systems Biology

Many of the most compelling and interesting examples of convergent evolution can be classified as complex and are likely due to interactions between many players (see examples above). Although the importance of studying these complex interactions is widely noted, the tools for studying them are not so widely used and often in need of improvement (Manolio et al., 2009; Zuk et al., 2012; Chandler et al., 2014; Taylor and Ehrenreich, 2014, 2015).

Systems biology generally denotes the study of biological systems, with multiple components and interactions (Kitano, 2002a, b; Loewe, 2009). The systems under study can vary greatly in scale, from whole ecosystems down to interactions between proteins. Systems biology offers a variety of developed and developing tools, such as network modeling and flux balance analysis, which are ideal for the study of complex convergent phenotypes because they have the ability to consider the many players and interactions involved in the phenotypes of interest.

Systems biology methods are also uniquely adapted to the discovery of emergent properties that are not apparent from studying single components of a system (the whole is more than the sum of its parts). Many complex traits can be considered emergent and are difficult (or impossible) to understand using purely reductive approaches. For example, genome scale metabolic network modeling of single-gene mutants in the model plant *Arabidopsis thaliana* suggests that even simple mutations can have effects on multiple metabolites, even when those metabolites' location in the metabolic network are distant from that of the mutant gene (Kim et al., 2015). Our own work has shown that glucosinolates, defensive compounds including mustard oils that deter insect predators, are very costly for the plant to metabolize; supporting their important evolutionary roll in plant herbivore defense (Bekaert et al., 2012). The costliness of glucosinolates seems logical, but becomes much clearer with systems biology analysis. Modeling the enzymatic pathways of C_4 photosynthesis (as described above) is another example.

Several studies outside of plants also demonstrate the usefulness of systems biology for understanding complex and/or convergent traits. For instance, predictive models of the human microbiome offer valuable insights into how microorganisms

interact with human health (Greenblum et al., 2013; Levy and Borenstein, 2013), and network modeling of neuronal genes in autism patients has reveled groups of genes which interact together and are associated with autism and other human diseases (Hormozdiari et al., 2015). These types of studies are becoming increasingly possible in plants as molecular and genomic resources advance.

Surprising Examples of Convergence: Crabtree and Warburg Effects

Although systems biology approaches are being increasingly used in plants, the resources available are still limited in comparison to those found in the human and microbial research communities. For this reason, methods and discoveries currently occurring in these communities are likely to provide a road map for future discoveries in plants. Along this vein, we share another human/microbe example that represents one of the best illustrations of how systems biology analyses can yield new and surprising insights. This example comes from the metabolic similarities observed among yeast, cancer cells, and embryos (Mayfield-Jones et al., 2013; Mordhorst et al., 2015). Most species of yeast will ferment sugars only in the absence of oxygen. However, bakers' yeast, the product of an ancient allopolyploidy event (Marcet-Houben and Gabaldón, 2015), commonly ferments sugars even when oxygen is present, a phenotype known as the Crabtree effect. This behavior is odd because such fermentation is apparently less energetically efficient than is the normal respiratory pathway of complete conversion of sugars to CO_2. The current explanation for this behavior relates to the tragedy of the commons (Hardin, 1968) and can be explained as follows. More efficient yeast cells

(using the normal respiratory pathway) are able to extract more chemical energy from ingested glucose, but the time required to do so means that they have reduced growth rates as compared to those using fermentation. For this reason, the fast, wasteful cells outcompete the efficient cells and dominate the culture (Pfeiffer et al., 2001; Pfeiffer and Schuster, 2005; MacLean and Gudelj, 2006; Dashko et al., 2014; Pfeiffer and Morley, 2014). Several researchers have suggested that the polyploidy event (or whole genome duplication) and the later preferential retention of glycolysis enzymes from that polyploidy may have been responsible for this shift in metabolic preference (Blank et al., 2005; Piškur et al., 2006; Conant and Wolfe, 2007; Merico et al., 2007) while others believe the Crabtree effect may pre-date the WGD event (Hagman et al., 2013).

What makes the Crabtree effect so interesting in terms of convergent evolution is that a similar phenomenon occurs in human cancer cells that are, of course, phylogenetically distant from yeast (Mayfield-Jones et al., 2013). In cancer cells, this phenomenon has been termed the Warburg effect and is again defined by glucose fermentation, in this case in tumor cells (Gatenby and Gillies, 2004). The Warburg effect often occurs when tumors are in a hypoxic condition and the citric acid cycle cannot be used for energy production. However, as with the Crabtree effect in yeast, the Warburg effect can also occur when oxygen is present, and the citric acid cycle would be more efficient (Gatenby and Gillies, 2004; Kim and Dang, 2006; Arora et al., 2015).

Whether or not the Crabtree and Warberg effects are convergent at some level is still unknown, but several lines of evidence suggest this possibility. For example, both the Crabtree and Warburg effects use similar mechanisms to repress oxidative phosphorylations by glucose (Diaz-Ruiz et al., 2011; Dell' Antone, 2012; Mayfield-Jones

et al., 2013). It is also plausible that both the Warburg and Crabtree effects may be the result of selection for fast but inefficient growth (Pfeiffer et al., 2001; Pfeiffer and Schuster, 2005; Mayfield-Jones et al., 2013). What makes this hypothesis even more interesting is the possible role of genome duplications or similar events playing a role in the both the Warburg and Crabtree effects (Blank et al., 2005; Piškur et al., 2006; Conant and Wolfe, 2007; Ganem et al., 2007; Merico et al., 2007; Merlo et al., 2010). The role of genome duplication in yeasts has already been mentioned: intriguingly tumor cells commonly also show large-scale copy number alterations that result from either whole-genome duplications or at least relatively large-scale aneuploidy (Shackney et al., 1989; Mitelman, 2000; Storchova and Pellman, 2004; Ganem et al., 2007; Fröhling and Döhner, 2008; Merlo et al., 2010).

Beyond applications to the study of cancer, the Crabtree and Warburg effects illustrate more generally the value of systems biology in understanding convergences. At a metabolic level, human cancer cells appear to be more similar to yeast cells then they are to the other (non-cancerous) human cells from which they originated. Without systems biology tools, who would have hypothesized that yeast and human cancer cells might share such a strong convergent phenotype? It is our hope and prediction that similar systems approaches will soon yield as unlikely and insightful results within plants as those described above for yeast and humans.

Doubling the System: Studying Convergent Gene Loss and Retention After Whole Genome Duplication

Another, related, example of convergent evolution, for which systems biology and comparative genomics tools have improved our understanding in plants and elsewhere, is the convergent pattern of genome evolution after polyploidy/whole genome duplications (WGD) events (Freeling, 2009;Bekaert et al., 2011; De Smet et al., 2013; Mayfield-Jones et al., 2013; Conant, 2014). These convergences fulfill the requirement of complexity described earlier, as the patterns observed often involve hundreds to thousands of genes and in some cases can be linked to relatively complex biological changes.

In order to understand the nature of these convergences, a brief detour into the biology of WGDs is needed. WGDs can occur in various ways but are most commonly caused by errors in meiosis or mitosis (Mayfield-Jones et al., 2013; Mason and Pires, 2015). These events may occur within a single species, yielding two or more copies of the same genome (called autopolyploidy), or together with the hybridization of two distinct species, forming a new species with genome copies from both parents (known as allopolyploidy).

In all cases, a WGD event results in multiple copies of most or all genes/alleles in the genome(s). The majority of these duplicate genes are lost soon after the WGD event (Scannell et al., 2007). However, the surviving duplicates tend to be drawn from similar types of genes across a broad range of taxa including plants, yeast, vertebrates, and paramecium (Seoighe and Wolfe, 1999; Blanc and Wolfe, 2004; Aury et al., 2006; Paterson et al., 2006; Schnable et al., 2009, 2011a, 2011b; Huminiecki and Heldin, 2010;

De Smet et al., 2013; McGrath et al., 2014; Thompson et al., 2015). For instance, transcription factors, ribosomal proteins and kinases all are more commonly retained in duplicate after WGD than expected (Freeling, 2009; Thompson et al., 2015). Moreover, it is not only the set of duplicates retained after WGD that shows evidence of convergence. The genes that are rarely or never seen to survive in duplicate are also very similar across diverse taxa: angiosperms and yeast show convergent early losses of genes involved in DNA repair and genes whose products are targeted to the mitochondria (De Smet et al., 2013; Conant, 2014;). While this example of convergent evolution might seem trivial, it is more complex than it appears (more than 10% of the genes in a given genome may be involved) and can have far-reaching consequences (Huminiecki and Heldin, 2010).

The reasons for these similarities in gene retention were not initially known, but there is now a substantial body of theory that explains the convergence of gene loss and retention after WGD events, namely the dosage balance hypothesis (DBH) (Papp et al., 2003; Freeling and Thomas, 2006; Birchler and Veitia, 2007, 2010, 2012; Edger and Pires, 2009; Conant et al., 2014; Thompson et al., 2015). This hypothesis posits that maintaining a stoichiometric balance in the dosage of some genes (i.e., keeping those genes in the same relative copy number) is important to an organism's fitness and therefore maintained by selection. The reasons that dosage changes alter fitness are still imperfectly understood, but generally relate to the need for the interactions between gene products to occur in certain proportions (Veitia and Birchler, 2015). This type of selection acts both on individual duplications, which are disfavored under these conditions, and on the duplicates produced by a WGD, where it is the loss of certain gene duplicates after WGD that is detrimental due to the disruption in dosage that results. This

preferential retention of some genes after WGD is thought to be, at least in part, responsible for the state of modern genomes that show signatures of ancient WGD events. Not only are these examples of convergent preferential retention after WGD abundant in nature, but it is also possible to re-create and study these events in the laboratory, particularly in plants (Kato and Birchler, 2006; Gaeta et al., 2007; Gaeta and Pires, 2010; Mestiri et al., 2010; Buggs et al., 2011; Xiong et al., 2011; Buggs et al., 2012; Washburn and Birchler, 2014; Yoo et al., 2014; Li et al., 2015a). Likewise, comparative genomics allow us to track these processes over the evolutionary history of the lineages involved, observing the actions of dosage balance in the process (Conant, 2014). Hence, many examples of WGD followed by similar patterns of gene loss and retention represent convergent systems where phylogenetic independence and complexity are both well demonstrated; leaving selection and constraint as the likely causes.

Why Study Convergence in Plants?

Different biological systems have distinct advantages and disadvantages for the study of convergent evolution. Bacterial, fungal, and viral systems are particularly amenable to experimental testing of hypotheses regarding convergence because of their short generation times and ease of laboratory manipulation. However, systems such as plants (both single and multi-cellular) represent fertile ground for the study of convergent evolution, particularly given the wide availability and cost effectiveness of various -omics based technologies.

Several convergent plant phenotypes have received considerable attention over the years, including the multiple origins of carnivorous, parasitic, and mycoheterotrophic plants (Albert et al., 1992; Smith, 1997; Cameron et al., 2002; Soltis et al., 2005; Barkman et al., 2007; Hashimoto et al., 2012; Wicke et al., 2014; Pavlovič and Saganová, 2015). Each of these convergent phenotypes appear, at least on the surface, to be complex traits whose evolution could be better understood using systems biology.

In addition to C_4 photosynthesis, which is described in detail above, CAM photosynthesis (which increases a plant's water use efficiency by only opening its stomata at night) has also evolved multiple times from C_3 ancestry (Yang et al., 2015). Like C_4, CAM appears to be a complex trait with a complicated evolutionary history. CAM plants have received less attention and investment than C_4 in recent years, but the resources available for studying them on a systems level are quickly becoming available (Yang et al., 2015).

One plant and animal phenomenon that has been studied a great deal but deserves even more research, particularly in light of convergence, is the "domestication syndrome." This syndrome postulates that domestication represents a long-term convergent evolution experiment for which many of the selective pressures can be confidently inferred (Lin et al., 2012; Fuller et al., 2014; Larson et al., 2014; Schmutz et al., 2014; Gaut, 2015; Takuno et al., 2015). Whether or not domestication should really be considered a syndrome or more of a process with many events has been questioned in recent literature (Doebley et al., 2006; Gerbault et al., 2014; Larson et al., 2014), but many domesticated crops do share similar traits that make them easier to use in agricultural settings. These traits include the loss of seed dispersal mechanisms, shorter

flowering times, and increased seed yields. Domesticated animals have also been selected for a suite of similar traits including tameness, changes in coat color, reduction in tooth size, and reduction in brain size (Diamond, 2002; Wilkins et al., 2014). Because the wild relatives of domesticated species often display very different phenotypes for these characters, it seems unlikely that they result from constraints on evolution alone, making them an interesting case for the study of selection in convergence.

A recent study looked for evidence of constraints governing which genes were selected for during the domestication of corn and rice (Gaut, 2015). If selection favored the same genes in both cases, it would suggest that constraint played a role in the process. The study was unable to find such evidence, leading to the conclusion that constraint does not play a major role (Gaut, 2015). Of course, as the author of the study acknowledged, it was only a preliminary comparison between two species and may not apply to domestication in general. Several other studies have in fact found evidence for gene reuse in domesticated species, implying a small suite of genes are available for alteration through artificial selection, and therefore some amount of constraint might be responsible for domestication-related phenotypes (Paterson et al., 1995; Ramsay et al., 2011; Butelli et al., 2012; Lenser and Theißen, 2013; Martin and Orgogozo, 2013). Further application of genomics and systems biology tools to the domestication syndrome will likely yield valuable insights into this phenomena.

Hybrid vigor or heterosis is another convergent phenomenon that takes place in a variety of organisms but is particularly common and well studied in plants (Birchler et al., 2010). Heterosis describes the phenomenon that, when two distantly related organisms or genotypes are crossed with each other, their progeny are often more

vigorous (in one aspect or another) than either of the two original parents. Various explanations have been given for this phenomenon, but a clear consensus on its cause(s) across organisms has yet to be reached (J ones, 1917; East, 1936; Crow, 1948; Shull, 1948; Lippman and Zamir, 2007; Birchler et al., 2010; Chen, 2010; Freeling et al., 2012; Washburn and Birchler, 2014). Furthermore, all current explanations for heterosis fail as a general description of the phenomenon (Birchler, 2013). In other words, examples of heterosis exist which cannot be well explained by any of the current hypotheses.

As stated earlier, the ubiquity of polyploidy in plants also makes them fertile ground for the study WGD-related convergences. There have already been decades of experimentation, theory, and descriptive work on plant polyploidy (Soltis et al., 2010; 2012; Stebbins, 1950), and a number of methods exist for the experimental induction of WGD events in plants (Eigsti, 1938; Eigsti et al., 1949; Kato and Geiger, 2002; Kato and Birchler, 2006; Lukens et al., 2006; Tate et al., 2006; Gaeta et al., 2007; Yu et al., 2009; Gaeta and Pires, 2010). The combination of deep knowledge of the natural history of plant polyploids and the ability to experimentally manipulate them returns us to Gould's metaphor, except that now his thought experiment becomes the potential for experiments that really do replay this tape in plants.

Conclusions

The study of convergent evolution has fascinated biologists since the advent of evolutionary theory. Systems biology tools, –omics level technologies, and a plethora of convergent phenotypes (found in nature and inducible in the lab) place plant biologists in

an unprecedented position to dissect and understand convergent evolution. Compelling examples such as metabolic similarities between yeast and cancer, and the patterns of preferential gene retention in polyploids demonstrate the power and utility of emerging systems biology tools. Domestication, polyploidy, photosynthetic mode, and heterosis are among the many examples of convergent phenotypes that make plant systems ideal for the study of convergent evolution. Each of these examples awaits the application of systems biology to illuminate it in new ways.

Literature Cited

Albert V, S Williams, M Chase 1992 Carnivorous plants: Phylogeny and structural evolution. Science 257: 1491-1495.

Arora R, D Schmitt, B Karanam, M Tan, C Yates, W Dean-Colomb 2015 Inhibition of the warburg effect with a natural compound reveals a novel measurement for determining the metastatic potential of breast cancers. Oncotarget 6: 662-678.

Aury J-M, O Jaillon, L Duret, B Noel, C Jubin, BM Porcel, B Segurens, V Daubin, V Anthouard, N Aiach, O Arnaiz, A Billaut, J Beisson, I Blanc, K Bouhouche, F Camara, S Duharcourt, R Guigo, D Gogendeau, M Katinka, A-M Keller, R Kissmehl, C Klotz, F Koll, A Le Mouel, G Lepere, S Malinsky, M Nowacki, JK Nowak, H Plattner, J Poulain, F Ruiz, V Serrano, M Zagulski, P Dessen, M Betermier, J Weissenbach, C Scarpelli, V Schachter, L Sperling, E Meyer, J Cohen, P Wincker 2006 Global trends of whole-genome duplications revealed by the ciliate paramecium tetraurelia. Nature 444: 171-178.

Bale R, ID Neveln, APS Bhalla, MA MacIver, NA Patankar 2015 Convergent evolution of mechanically optimal locomotion in aquatic invertebrates and vertebrates. PLoS Biol 13: e1002123.

Barkman T, J McNeal, S-H Lim, G Coat, H Croom, N Young, C dePamphilis 2007 Mitochondrial DNA suggests at least 11 origins of parasitism in angiosperms and reveals genomic chimerism in parasitic plants. BMC Evol Biol 7: 248.

Barrick JE, RE Lenski 2013 Genome dynamics during experimental evolution. Nat Rev Genet 14: 827-839.

Bekaert M, PP Edger, CM Hudson, JC Pires, GC Conant 2012 Metabolic and evolutionary costs of herbivory defense: Systems biology of glucosinolate synthesis. New Phytol 196: 596-605.

Bekaert M, PP Edger, JC Pires, GC Conant 2011 Two-phase resolution of polyploidy in the arabidopsis metabolic network gives rise to relative and absolute dosage constraints. Plant Cell 23: 1719-1728.

Birchler JA 2013 Genetic rules of heterosis in plants. Pages 313-321 *in* Polyploid and hybrid genomics. John Wiley & Sons, Inc.

Birchler JA, RA Veitia 2007 The gene balance hypothesis: From classical genetics to modern genomics. Plant Cell 19: 395-402.

Birchler JA, RA Veitia 2010 The gene balance hypothesis: Implications for gene regulation, quantitative traits and evolution. New Phytol 186: 54-62.

Birchler JA, RA Veitia 2012 Gene balance hypothesis: Connecting issues of dosage sensitivity across biological disciplines. Proc Natl Acad Sci U S A 109: 14746-14753.

Birchler JA, H Yao, S Chudalayandi, D Vaiman, RA Veitia 2010 Heterosis. Plant Cell 22: 2105-2112.

Blanc G, KH Wolfe 2004 Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. Plant Cell 16: 1667-1678.

Blank LM, F Lehmbeck, U Sauer 2005 Metabolic-flux and network analysis in fourteen hemiascomycetous yeasts.

Bräutigam A, S Schliesky, C Külahoglu, CP Osborne, APM Weber 2014 Towards an integrative model of c4 photosynthetic subtypes: Insights from comparative transcriptome analysis of nad-me, nadp-me, and pep-ck c4 species. J Exp Bot.

Buggs RJ, L Zhang, N Miles, JA Tate, L Gao, W Wei, PS Schnable, WB Barbazuk, PS Soltis, DE Soltis 2011 Transcriptomic shock generates evolutionary novelty in a newly formed, natural allopolyploid plant. Curr Biol 21: 551-556.

Buggs Richard JA, S Chamala, W Wu, Jennifer A Tate, Patrick S Schnable, Douglas E Soltis, Pamela S Soltis, WB Barbazuk 2012 Rapid, repeated, and clustered loss of duplicate genes in allopolyploid plant populations of independent origin. Curr Biol 22: 248-252.

Burke MK, JP Dunham, P Shahrestani, KR Thornton, MR Rose, AD Long 2010 Genome-wide analysis of a long-term evolution experiment with drosophila. Nature 467: 587-590.

Butelli E, C Licciardello, Y Zhang, J Liu, S Mackay, P Bailey, G Reforgiato-Recupero, C Martin 2012 Retrotransposons control fruit-specific, cold-dependent accumulation of anthocyanins in blood oranges. Plant Cell 24: 1242-1255.

Cameron KM, KJ Wurdack, RW Jobson 2002 Molecular evidence for the common origin of snap-traps among carnivorous plants. Am J Bot 89: 1503-1509.

Castoe TA, APJ de Koning, H-M Kim, W Gu, BP Noonan, G Naylor, ZJ Jiang, CL Parkinson, DD Pollock 2009 Evidence for an ancient adaptive episode of convergent molecular evolution. Proc Natl Acad Sci USA 106: 8986-8991.

Chalhoub B, F Denoeud, S Liu, IAP Parkin, H Tang, X Wang, J Chiquet, H Belcram, C Tong, B Samans, M Corréa, C Da Silva, J Just, C Falentin, CS Koh, I Le Clainche, M Bernard, P Bento, B Noel, K Labadie, A Alberti, M Charles, D Arnaud, H Guo, C Daviaud, S Alamery, K Jabbari, M Zhao, PP Edger, H Chelaifa, D Tack, G Lassalle, I Mestiri, N Schnel, M-C Le Paslier, G Fan, V Renault, PE Bayer, AA Golicz, S Manoli, T-H Lee, VHD Thi, S Chalabi, Q Hu, C Fan, R Tollenaere, Y Lu, C Battail, J Shen, CHD Sidebottom, X Wang, A Canaguier, A Chauveau, A Bérard, G Deniot, M Guan, Z Liu, F Sun, YP Lim, E Lyons, CD Town, I Bancroft, X Wang, J Meng, J Ma, JC Pires, GJ King, D Brunel, R Delourme, M Renard, J-M Aury, KL Adams, J Batley, RJ Snowdon, J Tost, D Edwards, Y Zhou, W Hua, AG Sharpe, AH Paterson, C Guan, P Wincker 2014 Early allopolyploid evolution in the post-neolithic brassica napus oilseed genome. Science 345: 950-953.

Chandler CH, S Chari, D Tack, I Dworkin 2014 Causes and consequences of genetic background effects illuminated by integrative genomic analysis. Genetics 196: 1321-1336.

Chen L, AL DeVries, C-HC Cheng 1997 Evolution of antifreeze glycoprotein gene from a trypsinogen gene in antarctic notothenioid fish. Proc Natl Acad Sci USA 94: 3811-3816.

Chen ZJ 2010 Molecular mechanisms of polyploidy and hybrid vigor. Trends Plant Sci 15: 57-71.

Christin P-A, G Besnard 2009 Two independent c4 origins in aristidoideae (poaceae) revealed by the recruitment of distinct phosphoenolpyruvate carboxylase genes. Am J Bot 96: 2234-2239.

Christin P-A, RP Freckleton, CP Osborne 2010a Can phylogenetics identify c4 origins and reversals? Trends Ecol Evol 25: 403-409.

Christin PA, DM Weinreich, G Besnard 2010b Causes and evolutionary significance of genetic convergence. Trends Genet 26: 400-405.

Conant GC 2014 Comparative genomics as a time machine: How relative gene dosage and metabolic requirements shaped the time-dependent resolution of yeast polyploidy. Mol Biol Evol 31: 3184-3193.

Conant GC, JA Birchler, JC Pires 2014 Dosage, duplication, and diploidization: Clarifying the interplay of multiple models for duplicate gene evolution over time. Curr Opin Plant Biol 19: 91-98.

Conant GC, A Wagner 2003 Convergent evolution of gene circuits. Nat Genet 34: 264-266.

Conant GC, KH Wolfe 2007 Increased glycolytic flux as an outcome of whole-genome duplication in yeast.

Conte GL, ME Arnegard, CL Peichel, D Schluter 2012 The probability of genetic parallelism and convergence in natural populations. Proc R Soc Lond [Biol] 279: 5039-5047.

Conway Morris S 2003 Life's solution : Inevitable humans in a lonely universe. Cambridge University Press, Cambridge, UK.

Copley SD 2003 Enzymes with extra talents: Moonlighting functions and catalytic promiscuity. Curr Opin Chem Biol 7: 265-272.

Crow JF 1948 Alternative hypotheses of hybrid vigor. Genetics 33: 477-487.

Cui H, D Kong, X Liu, Y Hao 2014 Scarecrow, scr-like 23 and short-root control bundle sheath cell fate and function in arabidopsis thaliana. Plant J 78: 319-327.

Currie A 2013 Convergence as evidence. Br J Philos Sci 64: 763-786.

Dashko S, N Zhou, C Compagno, J Piškur 2014 Why, when, and how did yeast evolve alcoholic fermentation? FEMS Yeast Res 14: 826-832.

de Oliveira Dal'Molin CG, LE Quek, RW Palfreyman, SM Brumbley, LK Nielsen 2010 Aragem, a genome-scale reconstruction of the primary metabolic network in arabidopsis. Plant Physiol 152: 579-589.

De Smet R, KL Adams, K Vandepoele, MCE Van Montagu, S Maere, Y Van de Peer 2013 Convergent gene loss following gene and genome duplications creates

single-copy families in flowering plants. Proc Natl Acad Sci USA 110: 2898-2903.

Delaux P-M, G Radhakrishnan, G Oldroyd 2015 Tracing the evolutionary path to nitrogen-fixing crops. Curr Opin Plant Biol 26: 95-99.

Dell' Antone P 2012 Energy metabolism in cancer cells: How to explain the warburg and crabtree effects? Medical Hypotheses 79: 388-392.

Denoeud F, L Carretero-Paulet, A Dereeper, G Droc, R Guyot, M Pietrella, C Zheng, A Alberti, F Anthony, G Aprea, J-M Aury, P Bento, M Bernard, S Bocs, C Campa, A Cenci, M-C Combes, D Crouzillat, C Da Silva, L Daddiego, F De Bellis, S Dussert, O Garsmeur, T Gayraud, V Guignon, K Jahn, V Jamilloux, T Joët, K Labadie, T Lan, J Leclercq, M Lepelley, T Leroy, L-T Li, P Librado, L Lopez, A Muñoz, B Noel, A Pallavicini, G Perrotta, V Poncet, D Pot, Priyono, M Rigoreau, M Rouard, J Rozas, C Tranchant-Dubreuil, R VanBuren, Q Zhang, AC Andrade, X Argout, B Bertrand, A de Kochko, G Graziosi, RJ Henry, Jayarama, R Ming, C Nagai, S Rounsley, D Sankoff, G Giuliano, VA Albert, P Wincker, P Lashermes 2014 The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. Science 345: 1181-1184.

Diamond J 2002 Evolution, consequences and future of plant and animal domestication. Nature 418: 700-707.

Diaz-Ruiz R, M Rigoulet, A Devin 2011 The warburg and crabtree effects: On the origin of cancer cell energy metabolism and of yeast glucose repression. Biochim Biophys Acta 1807: 568-576.

Doebley JF, BS Gaut, BD Smith 2006 The molecular genetics of crop domestication. Cell 127: 1309-1321.

Doyle JJ, MA Luckow 2003 The rest of the iceberg. Legume diversity and evolution in a phylogenetic context. Plant Physiol 131: 900-910.

East EM 1936 Heterosis. Genetics 21: 375-397.

Edger P, JC Pires 2009 Gene and genome duplications: The impact of dosage-sensitivity on the fate of nuclear genes. Chromosome Res 17: 699-717.

Eigsti OJ 1938 A cytological study of colchicine effects in the induction of polyploidy in plants. Proc Natl Acad Sci U S A 24: 56-63.

Eigsti OJ, P Dustin, Jr., N Gay-Winn 1949 On the discovery of the action of colchicine on mitosis in 1889. Science 110: 692.

Fisher RA 1934 Indeterminism and natural selection. Philosophy of Science 1: 99-117.

Fletcher GL, CL Hew, PL Davies 2001 Antifreeze proteins of teleost fishes. Annu Rev Physiol 63: 359.

Foote AD, Y Liu, GWC Thomas, T Vinar, J Alfoldi, J Deng, S Dugan, CE van Elk, ME Hunter, V Joshi, Z Khan, C Kovar, SL Lee, K Lindblad-Toh, A Mancia, R Nielsen, X Qin, J Qu, BJ Raney, N Vijay, JBW Wolf, MW Hahn, DM Muzny,

KC Worley, MTP Gilbert, RA Gibbs 2015 Convergent evolution of the genomes of marine mammals. Nat Genet 47: 272-275.

Fouracre JP, S Ando, JA Langdale 2014 Cracking the kranz enigma with systems biology. J Exp Bot.

Fragata I, P Simões, M Lopes-Cunha, M Lima, B Kellen, M Bárbaro, J Santos, MR Rose, M Santos, M Matos 2014 Laboratory selection quickly erases historical differentiation. PLoS ONE 9: e96227.

Freeling M 2009 Bias in plant gene content following different sorts of duplication: Tandem, whole-genome, segmental, or by transposition. Annu Rev Plant Biol 60: 433-453.

Freeling M, BC Thomas 2006 Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. Genome Res 16: 805-814.

Freeling M, MR Woodhouse, S Subramaniam, G Turco, D Lisch, JC Schnable 2012 Fractionation mutagenesis and similar consequences of mechanisms removing dispensable or less-expressed DNA in plants. Curr Opin Plant Biol 15: 131-139.

Fröhling S, H Döhner 2008 Chromosomal abnormalities in cancer. New Engl J Med 359: 722-734.

Fuller DQ, T Denham, M Arroyo-Kalin, L Lucas, CJ Stevens, L Qin, RG Allaby, MD Purugganan 2014 Convergent evolution and parallelism in plant domestication

revealed by an expanding archaeological record. Proc Natl Acad Sci USA 111: 6147-6152.

Futuyma DJ 1998 Evolutionary biology. Sinauer Associates Inc., Sunderland, MA.

Gaeta RT, JC Pires 2010 Homoeologous recombination in allopolyploids: The polyploid ratchet. New Phytol 186: 18-28.

Gaeta RT, JC Pires, F Iniguez-Luy, E Leon, TC Osborn 2007 Genomic changes in resynthesized brassica napus and their effect on gene expression and phenotype. Plant Cell 19: 3403-3417.

Ganem NJ, Z Storchova, D Pellman 2007 Tetraploidy, aneuploidy and cancer. Curr Opin Genet Dev 17: 157-162.

Gatenby RA, RJ Gillies 2004 Why do cancers have high aerobic glycolysis? Nat Rev Cancer 4: 891-899.

Gaut BS 2015 Evolution is an experiment: Assessing parallelism in crop domestication and experimental evolution: (nei lecture, smbe 2014, puerto rico). Mol Biol Evol 32: 1661-1671.

Gehring WJ, K Ikeo 1999 Pax 6: Mastering eye morphogenesis and eye evolution. Trends Genet 15: 371-377.

Gerbault P, RG Allaby, N Boivin, A Rudzinski, IM Grimaldi, JC Pires, C Climer Vigueira, K Dobney, KJ Gremillion, L Barton, M Arroyo-Kalin, MD Purugganan, R Rubio de Casas, R Bollongino, J Burger, DQ Fuller, DG Bradley, DJ Balding,

PJ Richerson, MTP Gilbert, G Larson, MG Thomas 2014 Storytelling and story testing in domestication. Proc Natl Acad Sci USA 111: 6159-6164.

Glover BJ, CA Airoldi, SF Brockington, M Fernández-Mazuecos, C Martínez-Pérez, G Mellers, E Moyroud, L Taylor 2015 How have advances in comparative floral development influenced our understanding of floral evolution? Int J Plant Sci 176: 307-323.

Gould SJ 1989 Wonderful life: The burgess shale and the nature of history. W.W. Norton, New York.

Grass Phylogeny Working Group II 2012 New grass phylogeny resolves deep evolutionary relationships and discovers c4 origins. New Phytol 193: 304-312.

Greenblum S, HC Chiu, R Levy, R Carr, E Borenstein 2013 Towards a predictive systems-level model of the human microbiome: Progress, challenges, and opportunities. Curr Opin Biotechnol 24: 810-820.

Hagman A, T Säll, C Compagno, J Piskur 2013 Yeast "make-accumulate-consume" life strategy evolved as a multi-step process that predates the whole genome duplication. PLoS ONE 8: e68734.

Halder G, P Callaerts, WJ Gehring 1995 Induction of ectopic eyes by targeting expression of the eyeless gene in drosophila. Science 267: 1788.

Hardin G 1968 The tragedy of the commons. Science 162: 1243-1248.

Hashimoto Y, S Fukukawa, A Kunishi, H Suga, F Richard, M Sauve, M-A Selosse 2012 Mycoheterotrophic germination of pyrola asarifolia dust seeds reveals convergences with germination in orchids. New Phytol 195: 620-630.

Heckmann D, S Schulze, A Denton, U Gowik, P Westhoff, Andreas PM Weber, Martin J Lercher Predicting c4 photosynthesis evolution: Modular, individually adaptive steps on a mount fuji fitness landscape. Cell 153: 1579-1588.

Hormozdiari F, O Penn, E Borenstein, EE Eichler 2015 The discovery of integrated gene networks for autism and related disorders. Genome Res 25: 142-154.

Huminiecki L, CH Heldin 2010 2r and remodeling of vertebrate signal transduction engine. BMC Biol 8: 146.

Huss-Danell K 1997 Tansley review no. 93. Actinorhizal symbioses and their n 2 fixation. New Phytol: 375-405.

Ibarra RU, JS Edwards, BO Palsson 2002 Escherichia coli k-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. Nature 420: 186-189.

Jeong H, SP Mason, AL Barabasi, ZN Oltvai 2001 Lethality and centrality in protein networks. Nature 411: 41-42.

Jones DF 1917 Dominance of linked factors as a means of accounting for heterosis. Genetics 2: 466-479.

Kato A, JA Birchler 2006 Induction of tetraploid derivatives of maize inbred lines by nitrous oxide gas treatment. J Hered 97: 39-44.

Kato A, HH Geiger 2002 Chromosome doubling of haploid maize seedlings using nitrous oxide gas at the flower primordial stage. Plant Breeding 121: 370-377.

Kawecki TJ, RE Lenski, D Ebert, B Hollis, I Olivieri, MC Whitlock 2012 Experimental evolution. Trends Ecol Evol 27: 547-560.

Kim J-w, CV Dang 2006 Cancer's molecular sweet tooth and the warburg effect. Cancer Res 66: 8927-8930.

Kim T, K Dreher, R Nilo-Poyanco, I Lee, O Fiehn, BM Lange, BJ Nikolau, L Sumner, R Welti, ES Wurtele, SY Rhee 2015 Patterns of metabolite changes identified from large-scale gene perturbations in arabidopsis using a genome-scale metabolic network. Plant Physiol 167: 1685-1698.

Kitano H 2002a Computational systems biology. Nature 420: 206-210.

Kitano H 2002b Systems biology: A brief overview. Science 295: 1662-1664.

Kornegay JR, JW Schilling, AC Wilson 1994 Molecular adaptation of a leaf-eating bird: Stomach lysozyme of the hoatzin. Mol Biol Evol 11: 921-928.

Kriener K, C O'HUigin, H Tichy, J Klein 2000 Convergent evolution of major histocompatibility complex molecules in humans and new world monkeys. Immunogenetics 51: 169-178.

Larson G, DR Piperno, RG Allaby, MD Purugganan, L Andersson, M Arroyo-Kalin, L Barton, C Climer Vigueira, T Denham, K Dobney, AN Doust, P Gepts, MTP Gilbert, KJ Gremillion, L Lucas, L Lukens, FB Marshall, KM Olsen, JC Pires, PJ

Richerson, R Rubio de Casas, OI Sanjur, MG Thomas, DQ Fuller 2014 Current perspectives and the future of domestication studies. Proc Natl Acad Sci USA 111: 6139-6146.

Lenser T, G Theißen 2013 Conservation of fruit dehiscence pathways between lepidium campestre and arabidopsis thaliana sheds light on the regulation of indehiscent. Plant J 76: 545-556.

Levy R, E Borenstein 2013 Metabolic modeling of species interaction in the human microbiome elucidates community-level assembly rules. Proc Natl Acad Sci U S A 110: 12804-12809.

Li A-l, S-f Geng, L-q Zhang, D-c Liu, L Mao 2015a Making the bread: Insights from newly synthesized allohexaploid wheat. Mol Plant 8: 847-859.

Li G, J Wang, SJ Rossiter, G Jones, JA Cotton, S Zhang 2008 The hearing gene prestin reunites echolocating bats. Proc Natl Acad Sci USA 105: 13959-13964.

Li H-L, W Wang, PE Mortimer, R-Q Li, D-Z Li, KD Hyde, J-C Xu, DE Soltis, Z-D Chen 2015b Large-scale phylogenetic analyses reveal multiple gains of actinorhizal nitrogen-fixing symbioses in angiosperms associated with climate change. Scientific Reports 5: 14023.

Lin Z, X Li, LM Shannon, C-T Yeh, ML Wang, G Bai, Z Peng, J Li, HN Trick, TE Clemente, J Doebley, PS Schnable, MR Tuinstra, TT Tesso, F White, J Yu 2012 Parallel domestication of the shattering1 genes in cereals. Nat Genet 44: 720-724.

Lippman ZB, D Zamir 2007 Heterosis: Revisiting the magic. Trends Genet 23: 60-66.

Liu Y, SJ Rossiter, X Han, JA Cotton, S Zhang 2010 Cetaceans on a molecular fast track to ultrasonic hearing. Curr Biol 20: 1834-1839.

Lobkovsky AE, EV Koonin 2012 Replaying the tape of life: Quantification of the predictability of evolution. Front Genet 3: 246.

Loewe L 2009 A framework for evolutionary systems biology. BMC Systems Biology 3: 27.

Losos JB 2011 Convergence, adaptation, and constraint. Evolution 65: 1827-1840.

Lukens LN, JC Pires, E Leon, R Vogelzang, L Oslach, T Osborn 2006 Patterns of sequence loss and cytosine methylation within a population of newly resynthesized brassica napus allopolyploids. Plant Physiol 140: 336-348.

MacLean RC, I Gudelj 2006 Resource competition and social conflict in experimental populations of yeast. Nature 441: 498-501.

Maier U-G, S Zauner, C Woehle, K Bolte, F Hempel, JF Allen, WF Martin 2013 Massively convergent evolution for ribosomal protein gene content in plastid and mitochondrial genomes. GBE 5: 2318-2329.

Mangan N, M Brenner 2014 Systems analysis of the co2 concentrating mechanism in cyanobacteria. eLife 3: e02043.

Manolio TA, FS Collins, NJ Cox, DB Goldstein, LA Hindorff, DJ Hunter, MI McCarthy, EM Ramos, LR Cardon, A Chakravarti, JH Cho, AE Guttmacher, A Kong, L

Kruglyak, E Mardis, CN Rotimi, M Slatkin, D Valle, AS Whittemore, M Boehnke, AG Clark, EE Eichler, G Gibson, JL Haines, TFC Mackay, SA McCarroll, PM Visscher 2009 Finding the missing heritability of complex diseases. Nature 461: 747-753.

Map of Life. 2015. Crystallins: Eye lens proteins.

Marcet-Houben M, T Gabaldón 2015 Beyond the whole-genome duplication: Phylogenetic evidence for an ancient interspecies hybridization in the baker's yeast lineage. PLoS Biol 13: e1002220.

Martin A, V Orgogozo 2013 The loci of repeated evolution: A catalog of genetic hotspots of phenotypic variation. Evolution 67: 1235-1250.

Mason AS, JC Pires 2015 Unreduced gametes: Meiotic mishap or evolutionary mechanism? Trends Genet 31: 5-10.

Matos M, P Simões, MA Santos, SG Seabra, GS Faria, F Vala, J Santos, I Fragata 2015 History, chance and selection during phenotypic and genomic experimental evolution: Replaying the tape of life at different levels. Front Genet 6.

Mayfield-Jones D, JD Washburn, T Arias, PP Edger, JC Pires, GC Conant 2013 Watching the grin fade: Tracing the effects of polyploidy on different evolutionary time scales. Semin Cell Dev Biol 24: 320-331.

McCutcheon JP, BR McDonald, NA Moran 2009 Convergent evolution of metabolic roles in bacterial co-symbionts of insects. Proc Natl Acad Sci USA 106: 15394-15399.

McCutcheon JP, NA Moran 2010 Functional convergence in reduced genomes of bacterial symbionts spanning 200 my of evolution. GBE 2: 708-718.

McGrath CL, J-F Gout, P Johri, TG Doak, M Lynch 2014 Differential retention and divergent resolution of duplicate genes following whole-genome duplication. Genome Res 24: 1665-1675.

Merico A, P Sulo, J Piškur, C Compagno 2007 Fermentative lifestyle in yeasts belonging to the saccharomyces complex. FEBS J 274: 976-989.

Merlo LMF, L Wang, JW Pepper, PS Rabinovitch, CC Maley 2010 Polyploidy, aneuploidy and the evolution of cancer. Pages 1-13 *in* RYC Poon ed. Polyplodization and cancer. Vol. 676. Springer New York.

Mestiri I, V Chagué, A-M Tanguy, C Huneau, V Huteau, H Belcram, O Coriton, B Chalhoub, J Jahier 2010 Newly synthesized wheat allohexaploids display progenitor-dependent meiotic stability and aneuploidy but structural genomic additivity. New Phytol 186: 86-101.

Meyer JR, DT Dobias, JS Weitz, JE Barrick, RT Quick, RE Lenski 2012 Repeatability and contingency in the evolution of a key innovation in phage lambda. Science 335: 428-432.

Miller BG, RT Raines 2004 Identifying latent enzyme activities: Substrate ambiguity within modern bacterial sugar kinases. Biochemistry 43: 6387-6392.

Miller BG, RT Raines 2005 Reconstitution of a defunct glycolytic pathway via recruitment of ambiguous sugar kinases. Biochemistry 44: 10776-10783.

Milo R, S Shen-Orr, S Itzkovitz, N Kashtan, D Chklovskii, U Alon 2002 Network motifs: Simple building blocks of complex networks. Science 298: 824-827.

Mitelman F 2000 Recurrent chromosome aberrations in cancer. MUTAT RES-REV MUTAT 462: 247-253.

Mordhorst B, M Wilson, G Conant 2015 Some assembly required: Evolutionary and systems perspectives on the mammalian reproductive system. Cell Tissue Res: 1-12.

Natarajan C, J Projecto-Garcia, H Moriyama, RE Weber, V Muñoz-Fuentes, AJ Green, C Kopuchian, PL Tubaro, L Alza, M Bulgarella, MM Smith, RE Wilson, A Fago, KG McCracken, JF Storz 2015 Convergent evolution of hemoglobin function in high-altitude andean waterfowl involves limited parallelism at the molecular sequence level. PLoS Genet 11: e1005681.

Nielsen R 2005 Molecular signatures of natural selection. Annu Rev Genet 39: 197-218.

O'Brien PJ, D Herschlag 1999 Catalytic promiscuity and the evolution of new enzymatic activities. Chem Biol 6: R91-R105.

Papp B, C Pal, LD Hurst 2003 Dosage sensitivity and the evolution of gene families in yeast. Nature 424: 194-197.

Parker J, G Tsagkogeorga, JA Cotton, Y Liu, P Provero, E Stupka, SJ Rossiter 2013 Genome-wide signatures of convergent evolution in echolocating mammals. Nature 502: 228-231.

Paterson AH, BA Chapman, JC Kissinger, JE Bowers, FA Feltus, JC Estill 2006 Many gene and domain families have convergent fates following independent whole-genome duplication events in arabidopsis, oryza, saccharomyces and tetraodon. Trends Genet 22: 597-602.

Paterson AH, Y-R Lin, Z Li, KF Schertz, JF Doebley, SRM Pinson, S-C Liu, JW Stansel, JE Irvine 1995 Convergent domestication of cereal crops by independent mutations at corresponding genetic loci. Science 269: 1714-1718.

Pavlovič A, M Saganová 2015 A novel insight into the cost–benefit model for the evolution of botanical carnivory. Ann Bot 115: 1075-1092.

Pearce T 2012 Convergence and parallelism in evolution: A neo-gouldian account. Br J Philos Sci 63: 429-448.

Pfeiffer T, A Morley 2014 An evolutionary perspective on the crabtree effect. Front Mol Biosci 1: 17.

Pfeiffer T, S Schuster 2005 Game-theoretical approaches to studying the evolution of biochemical systems. Trends Biochem Sci 30: 20-25.

Pfeiffer T, S Schuster, S Bonhoeffer 2001 Cooperation and competition in the evolution of atp-producing pathways. Science 292: 504-507.

Piškur J, E Rozpędowska, S Polakova, A Merico, C Compagno 2006 How did saccharomyces evolve to become a good brewer? Trends Genet 22: 183-186.

Preston JC, LC Hileman 2009 Developmental genetics of floral symmetry evolution. Trends Plant Sci 14: 147-154.

Prijambada ID, S Negoro, T Yomo, I Urabe 1995 Emergence of nylon oligomer degradation enzymes in pseudomonas aeruginosa pao through experimental evolution. Appl Environ Microbiol 61: 2020-2022.

Quiring R, U Walldorf, U Kloter, WJ Gehring 1994 Homology of the eyeless gene of drosophila to the small eye gene in mice and aniridia in humans. Science 265: 785-789.

Ramsay L, J Comadran, A Druka, DF Marshall, WTB Thomas, M Macaulay, K MacKenzie, C Simpson, J Fuller, N Bonar, PM Hayes, U Lundqvist, JD Franckowiak, TJ Close, GJ Muehlbauer, R Waugh 2011 Intermedium-c, a modifier of lateral spikelet fertility in barley, is an ortholog of the maize domestication gene teosinte branched 1. Nat Genet 43: 169-172.

Regel R, SR Matioli, WR Terra 1998 Molecular adaptation of drosophila melanogaster lysozymes to a digestive function. Insect Biochem Mol Biol 28: 309-319.

Roalson EH 2011 Origins and transitions in photosynthetic pathway types in monocots: A review and reanalysis. Pages 319-338 *in* AS Raghavendra, RF Sage eds. C4 photosynthesis and related co2 concentrating mechanisms. Springer, Dordrecht, the Netherlands.

Sage RF, L Meirong, RK Monson 1999 The taxonomic distribution of c4 photosynthesis. RF Sage, RK Monson eds. C4 plant biology. Academic Press, San Diego, CA, USA.

Sage RF, TL Sage, F Kocacinar 2012 Photorespiration and the evolution of c4 photosynthesis. Annu Rev Plant Biol 63: 19-47.

Scannell DR, AC Frank, GC Conant, KP Byrne, M Woolfit, KH Wolfe 2007 Independent sorting-out of thousands of duplicated gene pairs in two yeast species descended from a whole-genome duplication. Proc Natl Acad Sci USA 104: 8397-8402.

Schmutz J, PE McClean, S Mamidi, GA Wu, SB Cannon, J Grimwood, J Jenkins, S Shu, Q Song, C Chavarro, M Torres-Torres, V Geffroy, SM Moghaddam, D Gao, B Abernathy, K Barry, M Blair, MA Brick, M Chovatia, P Gepts, DM Goodstein, M Gonzales, U Hellsten, DL Hyten, G Jia, JD Kelly, D Kudrna, R Lee, MMS Richard, PN Miklas, JM Osorno, J Rodrigues, V Thareau, CA Urrea, M Wang, Y Yu, M Zhang, RA Wing, PB Cregan, DS Rokhsar, SA Jackson 2014 A reference genome for common bean and genome-wide analysis of dual domestications. Nat Genet 46: 707-713.

Schnable JC, BS Pedersen, S Subramaniam, M Freeling 2011a Dose-sensitivity, conserved non-coding sequences, and duplicate gene retention through multiple tetraploidies in the grasses. Front Plant Sci 2: 2.

Schnable JC, NM Springer, M Freeling 2011b Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. Proc Natl Acad Sci U S A 108: 4069-4074.

Schnable PS, D Ware, RS Fulton, JC Stein, F Wei, S Pasternak, C Liang, J Zhang, L Fulton, TA Graves, P Minx, AD Reily, L Courtney, SS Kruchowski, C Tomlinson, C Strong, K Delehaunty, C Fronick, B Courtney, SM Rock, E Belter, F Du, K Kim, RM Abbott, M Cotton, A Levy, P Marchetto, K Ochoa, SM Jackson, B Gillam, W Chen, L Yan, J Higginbotham, M Cardenas, J Waligorski, E Applebaum, L Phelps, J Falcone, K Kanchi, T Thane, A Scimone, N Thane, J Henke, T Wang, J Ruppert, N Shah, K Rotter, J Hodges, E Ingenthron, M Cordes, S Kohlberg, J Sgro, B Delgado, K Mead, A Chinwalla, S Leonard, K Crouse, K Collura, D Kudrna, J Currie, R He, A Angelova, S Rajasekar, T Mueller, R Lomeli, G Scara, A Ko, K Delaney, M Wissotski, G Lopez, D Campos, M Braidotti, E Ashley, W Golser, H Kim, S Lee, J Lin, Z Dujmic, W Kim, J Talag, A Zuccolo, C Fan, A Sebastian, M Kramer, L Spiegel, L Nascimento, T Zutavern, B Miller, C Ambroise, S Muller, W Spooner, A Narechania, L Ren, S Wei, S Kumari, B Faga, MJ Levy, L McMahan, P Van Buren, MW Vaughn, K Ying, C-T Yeh, SJ Emrich, Y Jia, A Kalyanaraman, A-P Hsia, WB Barbazuk, RS Baucom, TP Brutnell, NC Carpita, C Chaparro, J-M Chia, J-M Deragon, JC Estill, Y Fu, JA Jeddeloh, Y Han, H Lee, P Li, DR Lisch, S Liu, Z Liu, DH Nagel, MC

McCann, P SanMiguel, AM Myers, D Nettleton, J Nguyen, BW Penning, L Ponnala, KL Schneider, DC Schwartz, A Sharma, C Soderlund, NM Springer, Q Sun, H Wang, M Waterman, R Westerman, TK Wolfgruber, L Yang, Y Yu, L Zhang, S Zhou, Q Zhu, JL Bennetzen, RK Dawe, J Jiang, N Jiang, GG Presting, SR Wessler, S Aluru, RA Martienssen, SW Clifton, WR McCombie, RA Wing, RK Wilson 2009 The b73 maize genome: Complexity, diversity, and dynamics. Science 326: 1112-1115.

Schwab IR 2012 Evolution's witness: How eyes evolved. Oxford University Press, Inc., New York, New York, USA.

Seoighe C, KH Wolfe 1999 Yeast genome evolution in the post-genome era. Curr Opin Microbiol 2: 548-554.

Shackney SE, CA Smith, BW Miller, DR Burholt, K Murtha, HR Giles, DM Ketterer, AA Pollice 1989 Model for the genetic evolution of human solid tumors. Cancer Res 49: 3344-3354.

Shull GH 1948 What is "heterosis"? Genetics 33: 439-446.

Slewinski TL 2013 Using evolution as a guide to engineer kranz-type c4 photosynthesis. Front Plant Sci 4: 212.

Slewinski TL, AA Anderson, C Zhang, R Turgeon 2012 Scarecrow plays a role in establishing kranz anatomy in maize leaves. Plant Cell Physiol 53: 2030-2037.

Smith JF 1997 Molecular evolution and adaptive radiation in brocchinia (bromeliaceae: Pitcairnioideae) atop tepuis of the guayana shield. Pages 259-311 *in* TJ Givnish, KJ Sytsma eds. Molecular evolution and adaptive radiation. Cambridge University Press, New York.

Soltis DE, RJA Buggs, JJ Doyle, PS Soltis 2010 What we still don't know about polyploidy. Taxon 59: 1387-1403.

Soltis DE, PS Soltis, PK Endress, MW Chase 2005 Phylogeny and evolution of angiosperms. Sinauer Associates Incorporated.

Soltis DE, PS Soltis, DR Morgan, SM Swensen, BC Mullin, JM Dowd, PG Martin 1995 Chloroplast gene sequence data suggest a single origin of the predisposition for symbiotic nitrogen fixation in angiosperms. Proc Natl Acad Sci USA 92: 2647-2651.

Soltis PS, DE Soltis 1998 Molecular evolution of 18s rdna in angiosperms: Implications for character weighting in phylogenetic analysis. Pages 188-210 *in* DE Soltis, PS Soltis, JJ Doyle eds. Molecular systematics of plants ii. Springer US.

Soltis PS, DE Soltis 2012 Polyploidy and genome evolution. Springer-Verlag, Berlin Heidelberg.

Soltis PS, DE Soltis, PG Wolf, DL Nickrent, SM Chaw, RL Chapman 1999 The phylogeny of land plants inferred from 18s rdna sequences: Pushing the limits of rdna signal? Mol Biol Evol 16: 1774-1784.

Stebbins GL 1950 Variation and evolution in plants. Columbia University Press, New York.

Stern DL 2013 The genetic causes of convergent evolution. Nat Rev Genet 14: 751-764.

Stewart C-B, JW Schilling, AC Wilson 1987 Adaptive evolution in the stomach lysozymes of foregut fermenters. Nature 330: 401-404.

Storchova Z, D Pellman 2004 From polyploidy to aneuploidy, genome instability and cancer. Nat Rev Mol Cell Biol 5: 45-54.

Takuno S, P Ralph, K Swarts, RJ Elshire, JC Glaubitz, ES Buckler, MB Hufford, J Ross-Ibarra 2015 Independent molecular basis of convergent highland adaptation in maize. Genetics (Early Online August 1 , 2015).

Tate JA, Z Ni, A-C Scheen, J Koh, CA Gilbert, D Lefkowitz, ZJ Chen, PS Soltis, DE Soltis 2006 Evolution and expression of homeologous loci in tragopogon miscellus (asteraceae), a recent and reciprocally formed allopolyploid. Genetics 173: 1599-1611.

Taylor MB, IM Ehrenreich 2014 Genetic interactions involving five or more genes contribute to a complex trait in yeast. PLoS Genet 10: e1004324.

Taylor MB, IM Ehrenreich 2015 Higher-order genetic interactions and their contribution to complex traits. Trends Genet 31: 34-40.

Tenaillon O, A Rodríguez-Verdugo, RL Gaut, P McDonald, AF Bennett, AD Long, BS Gaut 2012 The molecular diversity of adaptive convergence. Science 335: 457-461.

Thompson A, HH Zakon, M Kirkpatrick 2015 Compensatory drift and the evolutionary dynamics of dosage-sensitive duplicate genes. Genetics (Early Online January 1 , 2015).

van Hoek MJA, P Hogeweg 2007 The role of mutational dynamics in genome shrinkage. Mol Biol Evol 24: 2485-2494.

Veitia R, J Birchler 2015 Models of buffering of dosage imbalances in protein complexes. Biology Direct 10: 42.

Wagner A, DA Fell 2001 The small world inside large metabolic networks. Proc R Soc Lond [Biol] 268: 1803-1810.

Wake DB, MH Wake, CD Specht 2011 Homoplasy: From detecting pattern to determining process and mechanism of evolution. Science 331: 1032-1035.

Wang Y, A Bräutigam, APM Weber, X-G Zhu 2014 Three distinct biochemical subtypes of c4 photosynthesis? A modelling analysis. J Exp Bot 65: 3567-3578.

Washburn JD, JA Birchler 2014 Polyploids as a “model system” for the study of heterosis. Plant Reprod 27: 1-5.

Washburn JD, JC Schnable, G Davidse, JC Pires 2015 Phylogeny and photosynthesis of the grass tribe paniceae. Am J Bot 102.

Watts DJ, SH Strogatz 1998 Collective dynamics of "small-world" networks. Nature 393: 440-442.

Weinreich DM, NF Delaney, MA DePristo, DL Hartl 2006 Darwinian evolution can follow only very few mutational paths to fitter proteins. Science 312: 111-114.

Wicke S, B Schäferhoff, CW dePamphilis, KF Müller 2014 Disproportional plastome-wide increase of substitution rates and relaxed purifying selection in genes of carnivorous lentibulariaceae. Mol Biol Evol 31: 529-545.

Wilkins AS, RW Wrangham, WT Fitch 2014 The “domestication syndrome” in mammals: A unified explanation based on neural crest cell behavior and genetics. Genetics 197: 795-808.

Xiong Z, RT Gaeta, JC Pires 2011 Homoeologous shuffling and chromosome compensation maintain genome balance in resynthesized allopolyploid brassica napus. Proc Natl Acad Sci USA 108: 7908-7913.

Yang X, JC Cushman, AM Borland, EJ Edwards, SD Wullschleger, GA Tuskan, NA Owen, H Griffiths, JAC Smith, HC De Paoli, DJ Weston, R Cottingham, J Hartwell, SC Davis, K Silvera, R Ming, K Schlauch, P Abraham, JR Stewart, H-B Guo, R Albion, J Ha, SD Lim, BWM Wone, WC Yim, T Garcia, JA Mayer, J Petereit, SS Nair, E Casey, RL Hettich, J Ceusters, P Ranjan, KJ Palla, H Yin, C Reyes-García, JL Andrade, L Freschi, JD Beltrán, LV Dever, SF Boxall, J Waller, J Davies, P Bupphada, N Kadu, K Winter, RF Sage, CN Aguilar, J Schmutz, J Jenkins, JAM Holtum 2015 A roadmap for research on crassulacean acid

metabolism (cam) to enhance sustainable food and bioenergy production in a hotter, drier world. New Phytol 207: 491-504.

Yoo M-J, X Liu, JC Pires, PS Soltis, DE Soltis 2014 Nonadditive gene expression in polyploids. Annu Rev Genet 48: 485-517.

Yu Z, K Haage, V Streit, A Gierl, R Torres Ruiz 2009 A large number of tetraploid arabidopsis thaliana lines, generated by a rapid strategy, reveal high stability of neo-tetraploids during consecutive generations. Theor Appl Genet 118: 1107-1119.

Zaman L, JR Meyer, S Devangam, DM Bryson, RE Lenski, C Ofria 2014 Coevolution drives the emergence of complex traits and promotes evolvability. PLoS Biol 12: e1002023.

Zhang J, S Kumar 1997 Detection of convergent and parallel evolution at the amino acid sequence level. Mol Biol Evol 14: 527-536.

Zuk O, E Hechter, SR Sunyaev, ES Lander 2012 The mystery of missing heritability: Genetic interactions create phantom heritability. Proc Natl Acad Sci USA 109: 1193-1198.

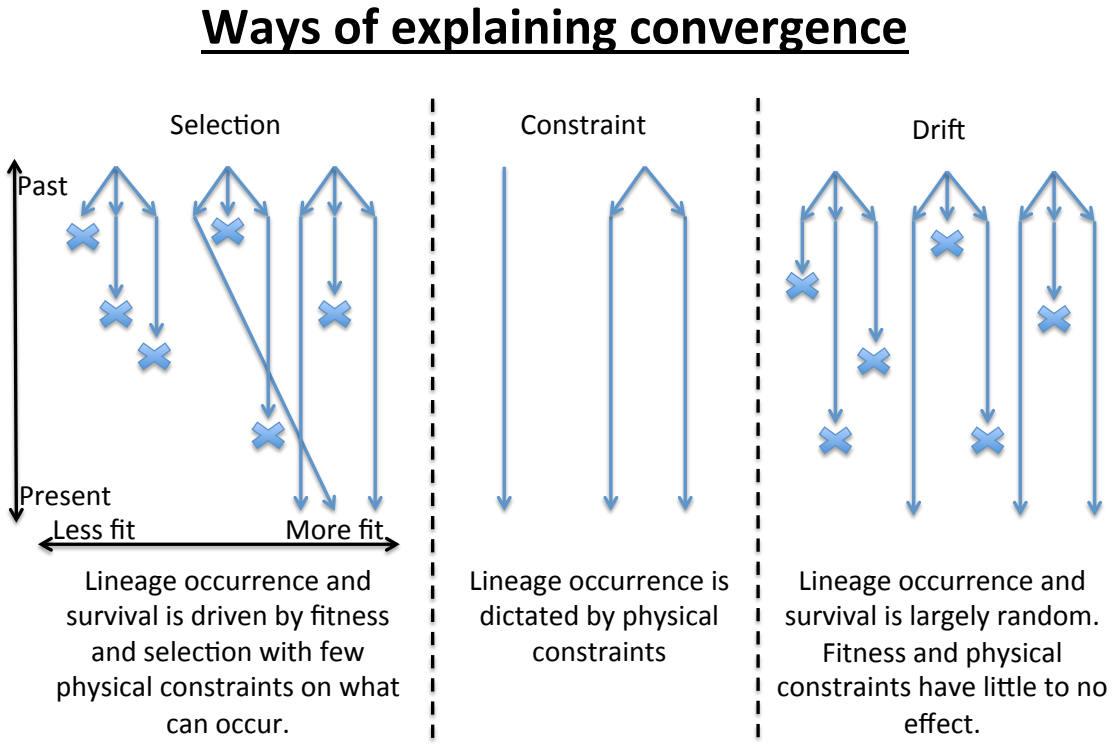


Figure 1.1 - Ways of explaining convergent evolution. Line drawings representing the evolutionary histories of distinct theoretical lineages with a convergent phenotype, and how these lineages might be expected to evolve under selection, constraint, or drift.

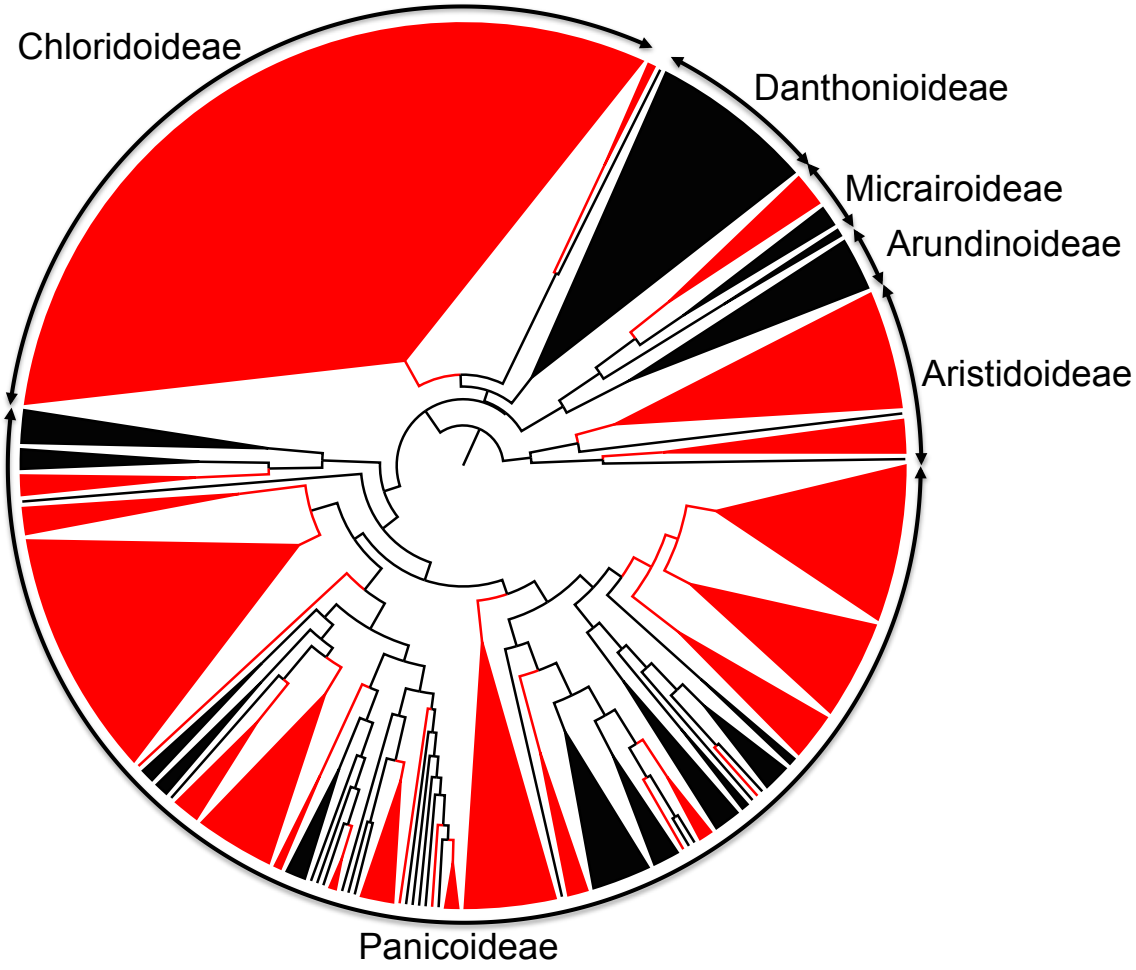


Figure 1.2 - Convergent evolution of C4 photosynthesis in the PACMAD grasses. A phylogenetic tree showing multiple origins of C_4 photosynthesis within a subsection of the grasses (family Poaceae) called the PACMAD. Lineages using the C_4 photosynthetic type are marked in red while those using the ancestral type (C_3) are marked in black. Note that the distribution of C_4 lineages shown here are not proportional to those found across angiosperms as a whole and are only an estimate of the number of origins within grasses. Several of these origins may be further subdivided or combined as future research demonstrates better support for distinct clades. The tree is re-drawn from Grass Phylogeny Working Group II, (2012) and Washburn et al., (2015).

CHAPTER 2: PHYLOGENY AND PHOTOSYNTHESIS OF THE GRASS TRIBE PANICEAE

Jacob D. Washburn[1], James C. Schnable[2], Gerrit Davidse[3], J. Chris Pires[1]

1 Division of Biological Sciences, University of Missouri, 311 Bond Life Sciences Center, Columbia, Missouri 65211 USA
2 Agronomy & Horticulture, University of Nebraska–Lincoln, Beadle Center E207, Lincoln, Nebraska 68583-0660 USA
3 Missouri Botanical Garden, P.O. Box 299, St. Louis, Missouri 63166-0299 USA

Please cite the published version in the American Journal of Botany:

Washburn JD, JC Schnable, G Davidse, JC Pires 2015 Phylogeny and photosynthesis of the grass tribe paniceae. Am J Bot 102: 1493-1505.
http://www.amjbot.org/content/102/9/1493.full

Abstract

PREMISE OF THE STUDY: The grass tribe Paniceae includes important food, forage, and bioenergy crops such as switchgrass, napiergrass, various millet species, and economically important weeds. Paniceae are also valuable for answering scientific and evolutionary questions about C_4 photosynthetic evolution, drought tolerance, and spikelet variation. However, the phylogeny of the tribe remains incompletely resolved.

METHODS: Forty-five taxa were selected from across the tribe Paniceae and outgroups for genome survey sequencing (GSS). These data were used to build a phylogenetic tree of the Paniceae based on 102 markers (78 chloroplast, 22 mitochondrial, 2 nrDNA). Ancestral state reconstruction analyses were also performed within the Paniceae using both the traditional and two subtype classification systems to test hypotheses of C_4 subtype evolution.

KEY RESULTS: The phylogenetic tree resolves many areas of the Paniceae with high support and provides insight into the origin and number of C_4 evolution events within the tribe. The recovered phylogeny and ancestral state reconstructions support between four and seven independent origins of C_4 photosynthesis within the tribe and indicate which species are potentially the closest C_3 sister taxa of each of these events.

CONCLUSIONS: Although the sequence of evolutionary events that produced multiple C_4 subtypes within the Paniceae remains undetermined, the results presented here are consistent with only a subset of currently proposed models. The species used in this study constitute a panel of C_3 and C_4 grasses that are suitable for further studies on C_4 photosynthesis, bioenergy, food and forage crops, and various developmental features of the Paniceae.

Introduction

The tribe Paniceae R.Br. s.s. (family Poaceae) includes 84 genera and ~1500 species of grasses (Morrone et al., 2012; Soreng et al., 2015). The tribe's members exist mainly in tropical and subtropical areas around the globe but also have limited distributions in cooler climates (Morrone et al., 2012). Paniceae have been circumscribed in various ways over the years, but are here referred to in the strict sense to include only the clade in Panicoideae with a base chromosome number of $x = 9$ (Grass Phylogeny Working Group II, 2012; Morrone et al., 2012; Soreng et al., 2015).

The tribe Paniceae are part of the subfamily Panicoideae and phylogenetically sister to the combined Andropogoneae and Paspaleae tribes that include economically important crops such as corn, sorghum, sugarcane, and a variety of turf and forage grasses. The tribe Paniceae itself contains many economically and socially important plants. For example, food crops found within the tribe include a number of "millet" grains, which are critical to human survival in the developing world and play important roles within the developed world as gluten-free alternatives to wheat and as domestic and agricultural animal feeds. Several members of Paniceae (i.e., switchgrass, guinea grass, elephant grass, and others) have also recently become the focus of breeding efforts as bioenergy crops. The tribe also includes important weed species within *Digitaria* and *Echinochloa*.

One suite of traits that makes the crop and weed species of Paniceae so ecologically successful is their drought tolerance and drought avoidance abilities, which have been linked to the use of the C_4 photosynthetic pathway, one of the most efficient ways plants turn sunlight into chemical energy (Sage et al., 2011). In fact, one member of

the Paniceae (*Panicum miliaceum,* known as common millet, broomcorn millet, or proso millet) has one of the lowest water requirements of any cultivated cereal although that may be due more to its rapid generation time than to drought tolerance (Baltensperger, 1996; Graybosch and Baltensperger, 2009; Hunt et al., 2014).

Currently, massive international efforts are focused on breeding and bioengineering C_4 photosynthesis into C_3 rice and other food and sustainable energy crops (Covshoff and Hibberd, 2012). However, C_4 comes in different shapes and sizes, and which C_4 subtype is most efficient for a given use will likely depend on the local environment, plant morphology, life history, and other traits. Over the past 30 million years, C_4 has evolved over 60 times (Sage et al., 2012). Because of this evolutionary history, C_4 is not "one" uniform photosynthetic type, but a diverse group of photosynthetic subtypes with over 20 different anatomies and three classically defined biochemical subtypes (Christin et al., 2015; Covshoff et al., 2014; Raghavendra, 1980; Sage, 2001; Sage, 2004; Sage et al., 2011). The tribe Paniceae also contains several clades where C_4 photosynthesis appears to have originated independently (Grass Phylogeny Working Group II, 2012).

The phylogeny of the tribe Paniceae has been investigated by various researchers over many years (Aliscioni et al., 2003; Bouchenak-Khelladi et al., 2009; Bouchenak-Khelladi et al., 2008; Bouchenak-Khelladi et al., 2010; Chemisquy et al., 2010; Christin et al., 2008; Christin et al., 2009a; Christin et al., 2009b; Christin et al., 2009c; Donadío et al., 2009; Doust and Kellogg, 2002; Duvall et al., 2007; Giussani et al., 2001; Grass Phylogeny Working Group, 2001; Grass Phylogeny Working Group II, 2012; Hodkinson et al., 2007; Morrone et al., 2012; Soreng et al., 2015; Spriggs et al., 2014;

Teerawatananon et al., 2011; Vicentini et al., 2008). The majority of these studies have focused on plastid markers (or nrDNA) and shown similar results but with poor backbone resolution. A few studies have included nuclear markers (other than nrDNA), and these have been incongruent with the plastid phylogenies(Christin et al., 2008; Christin et al., 2009a; Christin et al., 2007a; Christin et al., 2009b; Christin et al., 2009c; Christin et al., 2007b; Teerawatananon et al., 2011; Vicentini et al., 2008). This body of work was well summarized by a few recently published phylogenies (Grass Phylogeny Working Group II, 2012; Spriggs et al., 2014; Vicentini et al., 2008). The nuclear phylogeny by Vicentini et al. (2008) is well supported in terms of traditional phylogenetic statistics (maximum likelihood bootstraps and Bayesian posterior probability values) but is only based on two genes (one nuclear and one chloroplast). The plastid phylogeny is based on several chloroplast markers and a much more thorough species sampling of Paniceae taxa (Grass Phylogeny Working Group II, 2012; Spriggs et al., 2014). These two phylogenies (nuclear and chloroplast) disagree with each other in important areas of their topology, and the plastid phylogeny lacks the statistical support values necessary to resolve several key backbone areas of the tribe Paniceae (see Figure 2.1) (Grass Phylogeny Working Group II, 2012; Spriggs et al., 2014). Because of this lack of resolution and the lack of agreement between the different published phylogenies, it is impossible to confidently estimate the backbone phylogeny of the Paniceae from the current literature.

One unique aspect of the tribe Paniceae is that it is the only known group utilizing all three traditionally defined C_4 enzymatic subtypes (called NADP-ME, PCK, NAD-ME after a prominent enzyme in each pathway) without any C_3 taxa separating them phylogenetically (Sage et al., 2011). Modifications to this traditional subtype

classification system have been suggested to make it more biologically meaningful (see Discussion for further details and implications of these modifications).

The area of the Paniceae where these subtypes occur is referred to here as the MPC clade (for the subtribes Melinidinae, Panicinae, and Cenchrinae) as in the paper by the Grass Phylogeny Working Group II (2012). The clade has drawn interest for understanding C_4 subtype evolution for years, and several different approaches have been taken to unraveling its evolutionary history (Christin et al., 2009b; Christin et al., 2009c; Christin et al., 2007b; Grass Phylogeny Working Group II, 2012; Vicentini et al., 2008). A starting point for understanding the evolution of the MPC is a clear phylogenetic understanding of its history. Past studies have demonstrated that the MPC is in fact monophyletic, but the exact relationships between the three subclades in the MPC remain unclear (see Figure 2.1) (Grass Phylogeny Working Group II, 2012; Vicentini et al., 2008). There is also conflict in the literature as to whether an additional C_4 subtribe, Anthephorinae, should be included as a member of the clade (Grass Phylogeny Working Group II, 2012; Vicentini et al., 2008). The resolution of both of these uncertainties is critical to a clear understanding of the evolution of C_4 subtypes/C_4 diversity within the MPC clade. To better understand the evolutionary history of the tribe Paniceae and the C_4 subtypes found within it, this study analyzed 45 taxa from across the tribe Paniceae and outgroups using genome survey sequencing (GSS) (Steele et al., 2012), also known as genome skimming (Edger et al., 2014; Straub et al., 2012; Weitemier et al., 2015; Wysocki et al., 2014). Phylogenetic trees of the Paniceae were constructed based on various combinations of 102 markers (78 chloroplast, 22 mitochondrial, 2 nuclear ribosomal DNA). These trees resolve many areas of the Paniceae with high support and

provide insight into the number of C_4 events within the tribe, their closest C_3 relatives, and the evolutionary history of the MPC. Ancestral state reconstruction analyses were also performed within the Paniceae using both the traditional and two subtype classification systems to test hypotheses of C_4 subtype evolution, allowing the prioritization of hypotheses for future testing. Genome size data were also taken for members of the tribe to provide a resource for future studies within Paniceae.

Materials and methods

Taxon sampling

Sampling included 45 new taxa and five more from public archives for a total of 50 taxa from across the grasses (the main focus being within the tribe Paniceae). Most of the taxa were obtained from the USDA Germplasm system and grown in the greenhouses at the University of Missouri in Columbia, Missouri, United States. Others were received from various sources as noted in Supplemental Figure S2.1. Whenever possible, flowering specimens were collected for each taxon and evaluated against voucher specimens at the Missouri Botanical Garden Herbarium (MO). When significant differences between the specimens and vouchers were identified, those specimens were keyed out and again compared with vouchers until a confident match could be made. In several cases, the USDA specimens proved to be incorrectly identified to genus and/or species. The final identification of each specimen along with its original USDA ID is given in Supplemental Figure S2.1 and has been passed on to the USDA for correction of their records. Herbarium vouchers are deposited at MO with the exception of

Paraneurachne muelleri, which was sampled from a voucher at NY and a few species for which flowering specimens were unavailable.

DNA extraction and sequencing

Tissues were sampled from live materials (except in the case of a few dried specimens), and DNA was extracted using the DNeasy Plant Mini Kit (QIAGEN, Germantown, Maryland, USA). For samples in which the DNeasy kit did not provide adequate amounts of DNA, a second extraction was performed using an in-house urea-based DNA extraction protocol (see Supplemental Figure S2.2). Several samples were also extracted from herbarium specimens using a modification of the DNeasy Plant Mini Kit procedure with 600 µL Buffer AP1 and a 1-h incubation time rather than that recommended in the manual. Sequencing libraries were prepared by the MU sequencing core facility using the TruSeq DNA Sample Preparation Kit (Illumina, San Diego, California, USA) and sequenced in a multiplexed fashion with 24 samples per lane and 2×100bp chemistry on an Illumina Hiseq. Three of the samples (including the herbarium material sample) were made into libraries using the Nextera DNA Sample Preparation Kit (Illumina) because of its low DNA input requirements and sequenced under the same multiplexing and sequencing method as the others.

Phylogenetic analysis

Raw data were quality filtered and trimmed following standard procedures (Babraham Bioinformatics, 2015; Schmieder and Edwards, 2011). Each set of high-quality reads was then sorted into potentially overlapping files based on blast similarity to

a database of chloroplast, mitochondrial, or nrDNA sequences from members of the Paniceae and close relatives. These databases consisted of the gene regions only. The groups of reads from each individual taxon (three files per taxon, one chloroplast, one mitochondrial, and one nrDNA) were then assembled separately using the program SPAdes (Bankevich et al., 2012; Nurk et al., 2013; Wysocki et al., 2014) and annotated via several in-house scripts primarily utilizing stand alone BLAST (Camacho et al., 2009). Gene trees, species trees, and concatenated trees were constructed using the programs RAxML and ASTRAL and were created for each organelle group separately as well as all grouped together in one joint analysis (Mirarab et al., 2014; Stamatakis, 2014, 2006) (Supplemental Figures S2.3-S2.5). The RAxML trees were generated using partitioning by gene and a GTR GAMMA model. DNA sequence data were deposited in the Sequence Read Archive (SRA) of the National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov/sra) with the identification numbers noted in Supplemental Figure S2.1. Alignments and trees were deposited in Dryad (http://dx.doi.org/10.5061/dryad.92137).

Genome size estimates were obtained using flow cytometry at the Benaroya Research Institute at Virginia Mason in Seattle, Washington. These estimates were obtained using a standard protocol with some modifications (Arumuganathan and Earle, 1991). The 1C genome size estimates in this study are given as one-half of the value obtained by flow cytometry and are not corrected for ploidy level (see Supplemental Figure S2.1).

Ancestral state reconstruction

Each taxon in the phylogeny was identified as utilizing the C_3, C_4 NADP-ME, C_4 NAD-ME, or C_4 PCK photosynthetic type based on published literature (Brown, 1977; Ehleringer and Pearcy, 1983; Grass Phylogeny Working Group II, 2012; Gutierrez et al., 1974; Hattersley and Stone, 1986; Hattersley and Watson, 1992; Hattersley et al., 1982; Ibrahim et al., 2009; Lin et al., 1993; Liu and Osborne, 2015; Morrone et al., 2012; Prendergast et al., 1987; Sage et al., 1999; Vicentini et al., 2008). The literature on photosynthetic biochemical subtype is good for the tribe Paniceae, but there are still various species for which the subtypes have only been determined based on leaf anatomy and/or phylogenetic relatedness (Sage et al., 1999). Assigning subtype based on relatedness is particularly problematic within the genus *Alloteropsis* where, to the best of our knowledge, only one species has been confirmed biochemically (*Alloteropsis semialata* subsp. *semialata*) and a few others have been looked at anatomically. For this study, biochemical determination was used as the gold standard in all cases, and when that was not available, an anatomical determination and phylogenetic relatedness were used. We also determined the C_4 subtype for each species given the use of a two-subtype classification system; we simply used the enzyme with the highest activity (excluding PCK) for the subtype and then used phylogenetic relatedness to infer the subtypes of those that have not been examined biochemically.

The subtypes were mapped onto the chloroplast tree and used to infer the number of C_4 origins as well as the number of origins of each of the C_4 subtypes. Ancestral state reconstruction analyses were performed using only the Paniceae clade of the tree with one outgroup for rooting and the assumption of C_3 as the ancestral photosynthetic state (Grass Phylogeny Working Group II, 2012). Multiple ancestral state reconstruction

analyses were performed using the program BEAST 2 (Bouckaert et al., 2014) (for making the tree ultrametric and for Bayesian character state reconstruction) and the ace function for likelihood ancestral state reconstruction in the R package ape (Paradis et al., 2014; The R Foundation, 2015). Reconstructions were preformed using a three-state model (NADP-ME, NAD-ME, C_3) and a four-state model (NADP-ME, PCK, NAD-ME, C_3). The R package phyloch was also used for various phylogenetic manipulations (Heibl, 2013).

Because of the low number of species sampled in the phylogeny, a tree that included the chloroplast data generated here as well as data used in the Grass Phylogeny Working Group II (2012) paper was also constructed. Each gene was aligned individually, concatenated into one large matrix, and the phylogeny was inferred using RAxML. The resulting tree had an identical, and well supported, backbone to that from the tree with only the original data generated in this study, but included the much greater species sampling of the Grass Phylogeny Working Group II (2012) paper. Following the method of the Grass Phylogeny Working Group II (2012) paper, species from the tips of clades within the tribe Paniceae portion of the tree were trimmed to include proportions of the species and photosynthetic type in each clade that are as similar as possible to those known to exist, while still including putative C_4 origins (Morrone et al., 2012). This trimming was done based on the criteria that it did not change the basic (subtribe level) backbone of the phylogeny. The trees were also trimmed so as to leave as many different genera in the tree as possible (i.e., if one genus had five representative species and another only had one, then a species would be removed from the first genus rather than the second). Photosynthetic subtype of each species in the tree was again determined as

stated earlier. Reconstructions were preformed using similar models to those just described but included an "unknown" state. In this case, a four-state model (NADP-ME, NAD-ME, C_3, unknown) and a five-state model (NADP-ME, PCK, NAD-ME, C_3, unknown) were used. Because of the size of this tree and computational constraints, the tree was made ultrametric, and an ancestral state reconstruction was performed using the chronopl function with lambda = 0.1 and the ace function, respectively, in the R package ape (Paradis et al., 2014).

Results

Trees were first constructed using markers for each genome separately (Figure 2.2, and Supplemental Figures S2.3 and S2.4). The mitochondrial and nuclear ribosomal trees were poorly resolved, suggesting insufficient marker information was present in these data sets alone. A combined tree (Supplemental Figure S2.5) generated using all three data types exhibited a similar topology to the plastid-only tree but with small changes (both increases and decreases) to the support assigned to particular groupings. All subsequent analyses were conducted using the 78-locus chloroplast tree (see Figure 2.1, 2.2).

A key result of this tree is that the monophletic nature of the subtribes Melinidinae, Panicinae, and Cenchrinae (earlier referred to a the MPC clade) are all well supported with Cenchrinae and Melinidinae being sister to each other and then sister to Panicinae. This result agrees with previously published nuclear phylogenies and is much better supported in this study than the most recent chloroplast phylogeny (Grass Phylogeny Working Group II, 2012; Vicentini et al., 2008).

Based on this study's sampling, the sister group to the MPC comprises a clade which includes the C_3 subtribe *Dichantheliinae,* the C_3 genus *Sacciolepis* and the mixed photosynthetic type subtribe Neurachninae. This result is different from both the (Grass Phylogeny Working Group II, 2012) plastid phylogeny and the Vicentini et al. (2008) nuclear phylogeny. What is not clear in the likelihood phylogeny (BS of 79 but PP of 1) is whether the *Sacciolepis* and *Dichantheliinae* clades are sister to each other or whether one of them is more closely related to the MPC than the other. The (Grass Phylogeny Working Group II, 2012) phylogeny and other chloroplast phylogenies were also unable to resolve this due to poor support values, but the (Vicentini et al., 2008) phylogeny places *Sacciolepis* as a closer relative to the MPC than *Dichantheliinae*. Moving out on the tree, the Boivinellinae is placed as the next sister group to the rest with greater support than it previously had and the Anthephorinae is placed as the next sister to the rest of the Paniceae. This placement is in agreement with past chloroplast trees, but is entirely inconsistent with the combined chloroplast and nuclear tree, which places Anthephorinae as sister to the MPC clade (see discussion section for further elaboration on the implications of this placement) (Grass Phylogeny Working Group, 2001; Grass Phylogeny Working Group II, 2012; Vicentini et al., 2008). Within subtribes of the Paniceae, we also find that both the genera *Digitaria* and *Urochloa* are paraphyletic. In both cases, this is in agreement with the Grass Phylogeny Working Group II (2012) tree.

Ancestral state reconstruction analyses of the Paniceae were performed using both the traditional subtype classification and the two subtype classification systems. These models indicate several possible scenarios for the evolution of C_4 and its subtypes within the tribe (see Figure 2.3–2.6).

Discussion

Chloroplast, mitochondrial, and nuclear incongruence

The combined 102-gene phylogeny (chloroplast, mitochondrial, and nrDNA) agrees generally with the topology of the chloroplast-only phylogeny (see Figure 2.2). However, when compared with the chloroplast-only tree, the combined tree shows enhanced support in some areas and decreased support in others (Supplemental Figure S2.5) suggesting that the mitochondrial and/or nrDNA tree(s) are either adding noise to the complete tree or that they actually support different topologies than the chloroplast data in certain areas. Most enhancements to the tree's support values were found within the smaller clades representing the different Paniceae subtribes, while the decreased support values were found within the backbone of the tree and within the individual clades. These observations are consistent with differing nuclear and chloroplast genome histories surrounding the origin of the Paniceae. This possibility seems particularly likely when one considers that the published nuclear gene phylogenies of the Paniceae (Teerawatananon et al., 2011; Vicentini et al., 2008) produce results incompatible with either this study or previous phylogenies based on plastid markers (Aliscioni et al., 2003; Bouchenak-Khelladi et al., 2009; Bouchenak-Khelladi et al., 2008; Bouchenak-Khelladi et al., 2010; Chemisquy et al., 2010; Christin et al., 2008; Donadío et al., 2009; Doust and Kellogg, 2002; Duvall et al., 2007; Giussani et al., 2001; Grass Phylogeny Working Group, 2001; Grass Phylogeny Working Group II, 2012; Hodkinson et al., 2007;

Morrone et al., 2012; Soreng et al., 2015; Spriggs et al., 2014). However, the limited number of nuclear markers examined to date are insufficient to draw any conclusions with certainty. Resolving this incongruence will require a new study employing data from a wider range of nuclear genes.

Multiple origins of C_4 photosynthesis

Several lines of evidence suggest that C_3 to C_4 transitions are much more common than reversions (Christin et al., 2013; Christin et al., 2010b; Grass Phylogeny Working Group II, 2012). Phylogenetic methods have been used to access the likelihood of C_3 to C_4 transitions and reversions within the grasses on several occasions (Christin et al., 2013; Christin et al., 2010a; Christin et al., 2011; Christin et al., 2012a; Christin et al., 2014; Christin et al., 2012b; Christin et al., 2010c; Giussani et al., 2001; Grass Phylogeny Working Group II, 2012; Khoshravesh et al., 2012; Vicentini et al., 2008). One of the most exhaustive of these analyses comes from the Grass Phylogeny Working Group II (2012) paper, which included a very large and representative species sampling. Their analysis indicates that C_3 to C_4 transitions happen at a rate which is ~50 times that of reversions within the grasses. Other observations support a high ratio of gains to losses for this trait, such as anatomical and enzymatic differences between different monophyletic C_4 groups (Bräutigam et al., 2014; Brown, 1977; Wang et al., 2014). For this reason, all occurrences of sister clades of C_3 and C_4 species that have a maximum likelihood (ML) support value of 80 or higher have been scored as C_3 to C_4 transitions rather than C_4 to C_3 reversions within this manuscript. However, the possibility of reversions cannot be entirely rule out.

Using this system, one can identify either six or seven distinct origins of C_4 photosynthesis or four origins and two to three evolutionary conversions between C_4 subtypes (the exact number depends on which subtype classification system is used) (Figure 2.2). It is of note that the tree in Figure 2.2 actually shows a minimum of five origins, but taking the conservative approach of considering only well resolved branches with ML support greater than 80 leaves only four origins that are well supported. The lowest possible number (four origins) results from lumping all C_4 subtypes into one category, under the assumption that different subtypes do not originate de novo, but are in fact only modifications of an original C_4 event. The higher estimate of six to seven C_4 origins is based on the assumption or definition of each C_4 subtype as its own C_4 origin. The four C_4 origins one can confidently confirm in the tree are likely an underestimate of the true number within the Paniceae. Both the genus *Alloteropsis* and the subtribe Neurachninae have been reported to contain multiple origins of C_4 photosynthesis (Christin et al., 2012a; Christin et al., 2012b); however, each is represented by only a single exemplar species in this study.

C_4 subtypes in the tribe Paniceae

While the traditional classification system has treated the identity of the primary decarboxylating enzyme in a C_4 species as a qualitative trait, more recent work suggests many species use multiple decarboxylating enzymes in varying ratios (Furbank, 2011). Phosphoenolpyruvate carboxykinase (PCK) activity has also been observed to play a significant role in the carbon shuttles of a range of species traditionally classified as belonging to the NADP-malic enzyme (ME) or NAD-ME enzymatic subtypes (Furbank,

2011; Wang et al., 2014). Theoretical modeling and observational studies suggest that the PCK cycle provides the greatest efficiency in terms of quanta of light required per molecule of CO_2 fixed (Ehleringer and Pearcy, 1983; von Caemmerer and Furbank, 1999); however, a more recent modeling study demonstrated that utilizing a pure PCK cycle requires a larger percentage of total light energy absorption in bundle sheath cells than may be achievable with traditional Kranz anatomy (Wang et al., 2014), providing one possible explanation for the observation that even in species where PCK appears to act as the primary decarboxylation enzyme significant levels of NADP-ME or NAD-ME activity are often observed (Gutierrez et al., 1974; Lin et al., 1993; Prendergast et al., 1987). These observations have called into question the usefulness of the traditional classification system and have led to the suggestion that all C_4 species be classified as either NAD-ME or NADP-ME (Wang et al., 2014).

On the other hand, much of the evidence discussed has been developed employing data from model species/genera that, under the traditional definition, employ NADP-ME or NAD-ME (Brown et al., 2005; Brutnell et al., 2010). There are also a number of species where PCK enzyme activity is far higher, even by an order of magnitude, than that of NADP-ME or NAD-ME (Gutierrez et al., 1974; Lin et al., 1993; Prendergast et al., 1987). These high PCK species often cluster together in phylogenetic studies of the MPC clade at the exclusion of species with the other two traditional subtypes (Grass Phylogeny Working Group II, 2012; Vicentini et al., 2008). Molecular evolution studies have demonstrated parallel positive selection specific to clades where PCK decarboxylation predominates (Christin et al., 2009a). Liu and Osborne (2015) also

recently demonstrated physiological differences in response to drought between grass species where NAD-ME predominates and those where PCK predominates.

Because the exclusive use of one system or the other has not been fully accepted by the field and may not even be appropriate for all situations, we have used both the traditional and the two subtype systems in our analyses.

C_4 subtype evolution in the MPC clade of the Paniceae

Several models for the evolution of C_4 subtypes within the Paniceae are equally consistent with the results of this study (see Figure 2.6). The first is that an ancestral C_4 subtype originally evolved within the MPC (shown as NAD-ME in Figure 2.6, but the other subtypes are equally likely) and that the other two subtypes evolved from it (Christin et al., 2009b). We call this the one subtype hypothesis. A second hypothesis (here called the three subtype hypothesis) is that all three subtypes evolved together and were present at some level within the most recent common ancestor (MRCA). Then one subtype or another became primary within each of the different lineages over time. A third hypothesis, here referred to as the C_3 hypothesis, is that the distinct subtypes each evolved from a C_3 MRCA (most likely preadapted to C_4 evolution).

The ancestral state reconstructions performed here had various outcomes depending on the methods used and assumptions made. The addition of species to the phylogeny (see Figures 2.4 and 2.5 as compared with Figure 2.3) paints a clearer picture of C_4 subtype evolution within the Paniceae, and several broad conclusions can be drawn from these reconstructions (Figure 2.3–2.5). First, the reconstructions using the two and three subtype systems differ from each other.

Under the traditional three subtype system, the reconstructions are not consistent with the one subtype hypothesis above and in Figure 2.6. Namely, the probability of any one of the C_4 subtypes being dominant over the others in the ancestral state is low within the model. The other two hypotheses are more difficult to confirm or deny based on the results described here. If one takes the most likely value for each of the nodes as its actual ancestral state, then C_3 is clearly in the MRCA of the MPC, which would at least partially support the third or C_3 hypothesis of each subtype evolving independently from C_3 photosynthesis or some sort of C_3–C_4 intermediate. Also consistent with this hypothesis (though not necessarily inconsistent with some variations of the others) is a published phylogenetic analysis of the PCK gene in the MPC clade. This analysis showed that the Melinidinae clade gained/co-opted its PCK gene after its split from the other two clades (Christin et al., 2009a; Christin et al., 2009b).

The remaining hypothesis (labeled as the three subtype hypothesis in Figure 2.6) suggests that all three subtypes existed within the MRCA of the MPC and that each of the subtypes has since become primary in one of the clades. The use of the same PEPC gene by the whole MPC supports this hypothesis (Christin et al., 2007b), as does evidence that the primary enzymes of each subtype are functional at some level within each of the MPC clades (Gutierrez et al., 1974; Lin et al., 1993; Prendergast et al., 1987).

Under the two subtype system, the reconstructions are somewhat different than above (see Figures 2.3, 2.5). The probability of NAD-ME being the ancestral subtype is very high, and the likelihood of any of the other hypotheses is very low, not unexpected given the shift in the number of species utilizing NAD-ME under the two subtype classification system. These results suggest that photosynthesis within the MPC clade has

undergone qualitative changes between C_3, C_4-NAD-ME, and C_4-NADP-ME, as well as at least one large quantitative flip-flop between low PCK with either high NAD-ME or NADP-ME and high PCK with either low levels of NAD-ME or NADP-ME.

The different predictions made, depending on the classification system used, indicate the important role these systems play in our ability to understand C_4 subtype evolution. The differences also demonstrate the need for more quantitative data and modeling to understand C_4 diversity at a higher level than simple C_4 subtype classification.

Two avenues of investigation may be particularly helpful to the current understanding of C_4 evolution within the MPC and the broader Paniceae. First, the phylogeny of the group should be investigated using many nuclear genes. Once the relationship of Anthephorineae to the MPC is more clearly understood one should be able to more confidently reconstruct the ancestral state of the MPC. Second, the use of genomic and metabolic network ancestral state reconstructions should be applied to understanding how the MPC may have evolved (Conant, 2014). Variations on these approaches have shed light on the evolution of various traits within organisms across the tree of life and would allow one to consider the entire dynamic pathway of each subtype and its diversity rather than simply examining the primary enzyme. Further elaboration on previous network modeling approaches are also likely to be informative to these questions (Wang et al., 2014).

This study provides insights into the chloroplast evolutionary history of the Paniceae along with the estimated nuclear genome sizes of various species across the tribe. The chloroplast and nuclear phylogenies of the tribe appear incongruent suggesting

the need for further investigation of the nuclear phylogeny. The evolutionary paths that led to multiple C_4 origins and multiple C_4 subtypes within the Paniceae remain undetermined, but progress was made in narrowing the range of hypotheses under which future studies of this evolution within the Paniceae should be made.

In conclusion, comparative studies of C_4 photosynthesis have largely focused on a small number of crop species and wild models likely as a result of a lack of access to germplasm and the intractability of most wild species, which exhibit poor germination, slow growth, possibility of polyploidy, and a lack of pre-existing molecular biology data. The accessions used in this study are largely available from the USDA, represent multiple origins of multiple subtypes of C_4 photosynthesis, and are amenable to experimentation. We hope that the characterization and availability of this diversity panel will drive the accumulation of further genomic and phenotypic data and catalyze new research into photosynthesis, food and forage crops, and plant development within the Paniceae.

Acknowledgments

The authors thank E. A. Kellogg for taxon sampling advice, plant materials, and critical review of the manuscript. They also thank K. M. Devos, T. P. Brutnell, L. E. Bartley, A. J. Studer, J. R. Burkhalter, W. M. Whitten, and D. B. Lowry for plant materials; R. Douglas for the urea-based DNA extraction protocol; and R. Hall for the modified DNeasy extraction protocol. Thank you to two anonymous reviewers for their constructive comments on the manuscript. Funding for this work was provided by grants from the National Science Foundation (DEB Award no. 1501406), the University of

Missouri Research Board, the University of Missouri Mizzou Advantage, and the Sigma Xi Grants-in-Aid of Research Program.

Literature Cited

Aliscioni SS, LM Giussani, FO Zuloaga, EA Kellogg 2003 A molecular phylogeny of panicum (poaceae: Paniceae): Tests of monophyly and phylogenetic placement within the panicoideae. Am J Bot 90: 796-821.

Arumuganathan K, ED Earle 1991 Nuclear DNA content of some important plant species. Plant Molecular Biology Reporter 9: 208-218.

Babraham Bioinformatics. 2015. Fastqc a quality control tool for high throughput sequence data.

Baltensperger DD. 1996. Foxtail and proso millet. In: J Janick ed. Progress in New Crops. ASHS Press, Alexandria, Va.

Bankevich A, S Nurk, D Antipov, AA Gurevich, M Dvorkin, AS Kulikov, VM Lesin, SI Nikolenko, S Pham, AD Prjibelski, AV Pyshkin, AV Sirotkin, N Vyahhi, G Tesler, MA Alekseyev, PA Pevzner 2012 Spades: A new genome assembly algorithm and its applications to single-cell sequencing J Comput Biol 19: 455-477.

Bouchenak-Khelladi Y, G Anthony Verboom, TR Hodkinson, N Salamin, O Francois, G NÍ Chonghaile, V Savolainen 2009 The origins and diversification of c4 grasses and savanna-adapted ungulates. Global Change Biol 15: 2397-2417.

Bouchenak-Khelladi Y, N Salamin, V Savolainen, F Forest, Mvd Bank, MW Chase, TR Hodkinson 2008 Large multi-gene phylogenetic trees of the grasses (poaceae): Progress towards complete tribal and generic level sampling. Mol Phylogen Evol 47: 488-505.

Bouchenak-Khelladi Y, GA Verboom, V Savolainen, TR Hodkinson 2010 Biogeography of the grasses (poaceae): A phylogenetic approach to reveal evolutionary history in geographical space and geological time. Bot J Linn Soc 162: 543-557.

Bouckaert R, J Heled, D Kühnert, T Vaughan, C-H Wu, D Xie, MA Suchard, A Rambaut, AJ Drummond 2014 Beast 2: A software platform for bayesian evolutionary analysis. PLoS Comp Biol 10: e1003537.

Bräutigam A, S Schliesky, C Külahoglu, CP Osborne, APM Weber 2014 Towards an integrative model of c4 photosynthetic subtypes: Insights from comparative transcriptome analysis of nad-me, nadp-me, and pep-ck c4 species. J Exp Bot 65: 3579-3593.

Brown NJ, K Parsley, JM Hibberd 2005 The future of c4 research – maize, flaveria or cleome? Trends Plant Sci 10: 215-221.

Brown WV 1977 The kranz syndrome and its subtypes in grass systematics. Memoirs of the Torrey Botanical Club 23: 1-97.

Brutnell TP, L Wang, K Swartwood, A Goldschmidt, D Jackson, XG Zhu, E Kellogg, J Van Eck 2010 Setaria viridis: A model for c4 photosynthesis. Plant Cell 22: 2537-2544.

Camacho C, G Coulouris, V Avagyan, N Ma, J Papadopoulos, K Bealer, T Madden 2009 Blast+: Architecture and applications. BMC Bioinformatics 10: 421.

Chemisquy MA, LM Giussani, MA Scataglini, EA Kellogg, O Morrone 2010 Phylogenetic studies favour the unification of pennisetum, cenchrus and odontelytrum (poaceae): A combined nuclear, plastid and morphological analysis, and nomenclatural combinations in cenchrus. Ann Bot 106: 107-130.

Christin P-A, G Besnard, E Samaritani, MR Duvall, TR Hodkinson, V Savolainen, N Salamin 2008 Oligocene co2 decline promoted c4 photosynthesis in grasses. Curr Biol 18: 37-43.

Christin P-A, SF Boxall, R Gregory, EJ Edwards, J Hartwell, CP Osborne 2013 Parallel recruitment of multiple genes into c4 photosynthesis. GBE 5: 2174-2187.

Christin P-A, RP Freckleton, CP Osborne 2010a Can phylogenetics identify c4 origins and reversals? Trends Ecol Evol 25: 403-409.

Christin P-A, B Petitpierre, N Salamin, L Büchi, G Besnard 2009a Evolution of c4 phosphoenolpyruvate carboxykinase in grasses, from genotype to phenotype. Mol Biol Evol 26: 357-365.

Christin P-A, TL Sage, EJ Edwards, RM Ogburn, R Khoshravesh, RF Sage 2011 Complex evolutionary transitions and the significance of c3-c4 intermediate forms of photosynthesis in molluginaceae. Evolution 65: 643-660.

Christin P-A, N Salamin, V Savolainen, MR Duvall, G Besnard 2007a C4 photosynthesis evolved in grasses via parallel adaptive genetic changes. Curr Biol 17: 1241-1247.

Christin P-A, E Samaritani, B Petitpierre, N Salamin, G Besnard 2009b Evolutionary insights on c4 photosynthetic subtypes in grasses from genomics and phylogenetics. GBE 1: 221-230.

Christin PA, M Arakaki, CP Osborne, EJ Edwards 2015 Genetic enablers underlying the clustered evolutionary origins of c4 photosynthesis in angiosperms. Mol Biol Evol 32: 846-858.

Christin PA, EJ Edwards, G Besnard, SF Boxall, R Gregory, EA Kellogg, J Hartwell, CP Osborne 2012a Adaptive evolution of c4 photosynthesis through recurrent lateral gene transfer. Curr Biol 22: 445-449.

Christin PA, RP Freckleton, CP Osborne 2010b Can phylogenetics identify c4 origins and reversals? Trends Ecol Evol 25: 403-409.

Christin PA, N Salamin, EA Kellogg, A Vicentini, G Besnard 2009c Integrating phylogeny into studies of c4 variation in the grasses. Plant Physiol 149: 82-87.

Christin PA, N Salamin, V Savolainen, G Besnard 2007b A phylogenetic study of the phosphoenolpyruvate carboxylase multigene family in poaceae: Understanding the molecular changes linked to c4 photosynthesis evolution. Kew Bulletin 62: 455-462.

Christin PA, E Spriggs, CP Osborne, CA Stromberg, N Salamin, EJ Edwards 2014 Molecular dating, evolutionary rates, and the age of the grasses. Syst Biol 63: 153-165.

Christin PA, MJ Wallace, H Clayton, EJ Edwards, RT Furbank, PW Hattersley, RF Sage, TD MacFarlane, M Ludwig 2012b Multiple photosynthetic transitions, polyploidy, and lateral gene transfer in the grass subtribe neurachninae. J Exp Bot 63: 6297-6308.

Christin PA, DM Weinreich, G Besnard 2010c Causes and evolutionary significance of genetic convergence. Trends Genet 26: 400-405.

Conant GC 2014 Comparative genomics as a time machine: How relative gene dosage and metabolic requirements shaped the time-dependent resolution of yeast polyploidy. Mol Biol Evol 31: 3184-3193.

Covshoff S, SJ Burgess, J Knerova, BM Kumpers 2014 Getting the most out of natural variation in c4 photosynthesis. Photosynthesis Res 119: 157-167.

Covshoff S, JM Hibberd 2012 Integrating c4 photosynthesis into c3 crops to increase yield potential. Curr Opin Biotechnol 23: 209-214.

Donadío S, LM Giussani, EA Kellogg, FO Zuolaga, O Morrone 2009 A preliminary molecular phylogeny of pennisetum and cenchrus (poaceae-paniceae) based on the trnl-f, rpl16 chloroplast markers. Taxon 58: 392-404.

Doust AN, EA Kellogg 2002 Inflorescence diversification in the panicoid “bristle grass” clade (paniceae, poaceae): Evidence from molecular phylogenies and developmental morphology. Am J Bot 89: 1203-1222.

Duvall MR, JI Davis, LG Clark, JD Noll, DH Goldman, JG Sánchez-Ken 2007 Phylogeny of the grasses (poaceae) revisited. Aliso: A Journal of Systematic and Evolutionary Botany 23: 237-247.

Edger PP, M Tang, KA Bird, DR Mayfield, G Conant, K Mummenhoff, MA Koch, JC Pires 2014 Secondary structure analyses of the nuclear rrna internal transcribed spacers and sssessment of its phylogenetic utility across the brassicaceae (mustards). PLoS ONE 9: e101341.

Ehleringer J, RW Pearcy 1983 Variation in quantum yield for co2 uptake among c3 and c4 plants. Plant Physiol 73: 555-559.

Furbank RT 2011 Evolution of the c4 photosynthetic mechanism: Are there really three c4 acid decarboxylation types? J Exp Bot 62: 3103-3108.

Giussani LM, JH Cota-Sánchez, FO Zuloaga, EA Kellogg 2001 A molecular phylogeny of the grass subfamily panicoideae (poaceae) shows multiple origins of c4 photosynthesis. Am J Bot 88: 1993-2012.

Grass Phylogeny Working Group 2001 Phylogeny and subfamilial classification of the grasses (poaceae). Annals of the Missouri Botanical Garden 88: 373-457.

Grass Phylogeny Working Group II 2012 New grass phylogeny resolves deep evolutionary relationships and discovers c4 origins. New Phytol 193: 304-312.

Graybosch RA, DD Baltensperger 2009 Evaluation of the waxy endosperm trait in proso millet (*panicum miliaceum*). Plant Breeding 128: 70-73.

Gutierrez M, VE Gracen, GE Edwards 1974 Biochemical and cytological relationships in c4 plants. Planta 119: 279-300.

Hattersley P, N Stone 1986 Photosynthetic enzyme activities in the c3-c4 intermediate *neurachne minor* s. T . Blake (poaceae). Funct Plant Biol 13: 399-408.

Hattersley PW, L Watson 1992 Diversification of photosynthesis. Pages 38-116 *in* GP Chapman ed. Grass evolution and domestication. Cambridge University Press, New York, NY, USA.

Hattersley PW, L Watson, CR Johnston 1982 Remarkable leaf anatomical variations in *neurachne* and its allies (poaceae) in relation to c3 and c4 photosynthesis. Bot J Linn Soc 84: 265-272.

Heibl C. 2013. Phyloch: Interfaces and graphic tools for phylogenetic data in r.

Hodkinson TR, N Salamin, MW Chase, Y Bouchenak-Khelladi, SA Renvoize, V Savolainen 2007 Large trees, supertrees, and diversification of the grass family. Aliso: A Journal of Systematic and Evolutionary Botany 23: 248-258.

Hunt HV, F Badakshi, O Romanova, CJ Howe, MK Jones, JSP Heslop-Harrison 2014 Reticulate evolution in *panicum* (poaceae): The origin of tetraploid broomcorn millet, *p. Miliaceum*. J Exp Bot 65: 3165-3175.

Ibrahim DG, T Burke, BS Ripley, CP Osborne 2009 A molecular phylogeny of the genus *alloteropsis* (panicoideae, poaceae) suggests an evolutionary reversion from c(4) to c(3) photosynthesis. Ann Bot 103: 127-136.

Khoshravesh R, A Hossein, TL Sage, B Nordenstam, RF Sage 2012 Phylogeny and photosynthetic pathway distribution in *anticharis* endl. (scrophulariaceae). J Exp Bot 63: 5645-5658.

Lin C, Y Tai, D Liu, M Ku 1993 Photosynthetic mechanisms of weeds in taiwan. Funct Plant Biol 20: 757-769.

Liu H, CP Osborne 2015 Water relations traits of c4 grasses depend on phylogenetic lineage, photosynthetic pathway, and habitat water availability. J Exp Bot 66: 761-773.

Mirarab S, R Reaz, MS Bayzid, T Zimmermann, MS Swenson, T Warnow 2014 Astral: Genome-scale coalescent-based species tree estimation. Bioinformatics 30: i541-i548.

Morrone O, L Aagesen, MA Scataglini, DL Salariato, SS Denham, MA Chemisquy, SM Sede, LM Giussani, EA Kellogg, FO Zuloaga 2012 Phylogeny of the paniceae (poaceae: Panicoideae): Integrating plastid DNA sequences and morphology into a new classification. Cladistics 28: 333-356.

Nurk S, A Bankevich, D Antipov, A Gurevich, A Korobeynikov, A Lapidus, A Prjibelsky, A Pyshkin, A Sirotkin, Y Sirotkin, R Stepanauskas, J McLean, R Lasken, S Clingenpeel, T Woyke, G Tesler, M Alekseyev, P Pevzner 2013 Assembling genomes and mini-metagenomes from highly chimeric reads. Pages 158-170 *in* M Deng, R Jiang, F Sun, X Zhang eds. Research in computational molecular biology. Vol. 7821. Lecture notes in computer science. Springer Berlin Heidelberg.

Paradis E, B Bolker, J Claude, HS Cuong, R Desper, B Durand, J Dutheil, O Gascuel, C Heibl, D Lawson, V Lefort, P Legendre, J Lemon, Y Noel, J Nylander, R Opgen-Rhein, A-A Popescu, K Schliep, K Strimmer, Dd Vienne. 2014. Ape: Analyses of phylogenetics and evolution.

Prendergast H, P Hattersley, N Stone 1987 New structural/biochemical associations in leaf blades of c4 grasses (poaceae). Funct Plant Biol 14: 403-420.

Raghavendra AS 1980 Characteristics of plant species intermediate between c3 and c4 pathways of photosynthesis: Their focus of mechanism and evolution of c4 syndrome. Photosynthetica 14: 271-273.

Sage RF 2001 Environmental and evolutionary preconditions for the origin and diversification of the c4 photosynthetic syndrome. Plant Biol 3: 202-213.

Sage RF 2004 The evolution of c4 photosynthesis. New Phytol 161: 341-370.

Sage RF, PA Christin, EJ Edwards 2011 The c4 plant lineages of planet earth. J Exp Bot 62: 3155-3169.

Sage RF, L Meirong, RK Monson 1999 The taxonomic distribution of c4 photosynthesis. Pages 551-584 *in* RF Sage, RK Monson eds. C4 plant biology. Academic Press, San Diego, CA, USA.

Sage RF, TL Sage, F Kocacinar 2012 Photorespiration and the evolution of c4 photosynthesis. Annu Rev Plant Biol 63: 19-47.

Schmieder R, R Edwards 2011 Quality control and preprocessing of metagenomic datasets. Bioinformatics 27: 863-864.

Soreng RJ, PM Peterson, K Romaschenko, G Davidse, FO Zuloaga, EJ Judziewicz, TS Filgueiras, JI Davis, O Morrone 2015 A worldwide phylogenetic classification of the poaceae (gramineae). J Syst Evol 53: 117-137.

Spriggs EL, P-A Christin, EJ Edwards 2014 C4 photosynthesis promoted species diversification during the miocene grassland expansion. PLoS ONE 9: e97722.

Stamatakis A 2006 Raxml-vi-hpc: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22: 2688-2690.

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

Steele PR, KL Hertweck, D Mayfield, MR McKain, J Leebens-Mack, JC Pires 2012 Quality and quantity of data recovered from massively parallel sequencing: Examples in asparagales and poaceae. Am J Bot 99: 330-348.

Straub SCK, M Parks, K Weitemier, M Fishbein, RC Cronn, A Liston 2012 Navigating the tip of the genomic iceberg: Next-generation sequencing for plant systematics. Am J Bot 99: 349-364.

Teerawatananon A, SWL Jacobs, TR Hodkinson 2011 Phylogenetics of panicoideae (poaceae) based on chloroplast and nuclear DNA sequences. Telopea 13: 115-142.

The R Foundation. 2015. R: The r project for statistical computing. The R Foundation.

Vicentini A, JC Barber, SS Aliscioni, LM Giussani, EA Kellogg 2008 The age of the grasses and clusters of origins of c4 photosynthesis. Global Change Biol 14: 2963-2977.

von Caemmerer S, RT Furbank 1999 Modeling c4 photosynthesis. RF Sage, RK Monson eds. C4 plant biology. Academic Press, San Diego.

Wang Y, A Bräutigam, APM Weber, X-G Zhu 2014 Three distinct biochemical subtypes of c4 photosynthesis? A modelling analysis. J Exp Bot 65: 3567-3578.

Weitemier K, SC Straub, M Fishbein, A Liston 2015 Intragenomic polymorphisms among high-copy loci: A genus-wide study of nuclear ribosomal DNA in *asclepias* (apocynaceae). PeerJ 3: e718.

Wysocki WP, LG Clark, SA Kelchner, SV Burke, JC Pires, PP Edger, DR Mayfield, JK Triplett, JT Columbus, AL Ingram, MR Duvall 2014 A multi-step comparison of short-read full plastome sequence assembly methods in grasses. Taxon 63: 899-910.

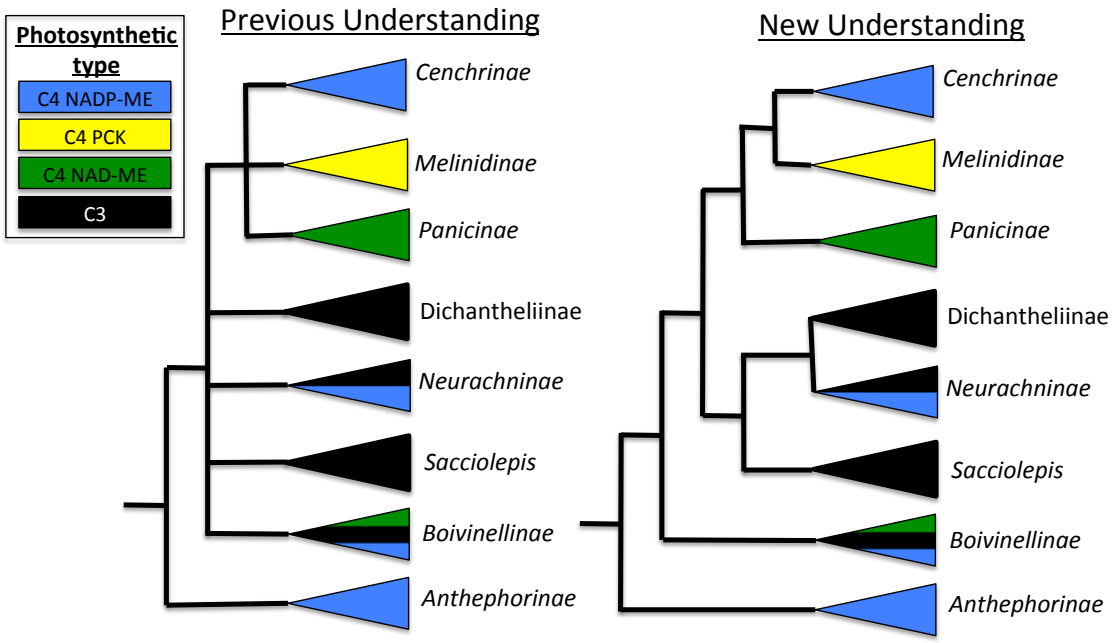


Figure 2.1 - Phylogenetic understandings of the tribe Paniceae. Left: Previous phylogeny of the tribe Paniceae redrawn from the Grass Phylogeny Working Group II (2012). Right: New phylogeny from the 78 chloroplast gene data set presented in this study. Branches with a posterior probability of less than 0.8 are collapsed into polytomies on both phylogenies. Figure is drawn using the traditional three C_4 sub-type system but can be easily interpreted based on the two sub-type system by replacing PCK with NAD-ME.

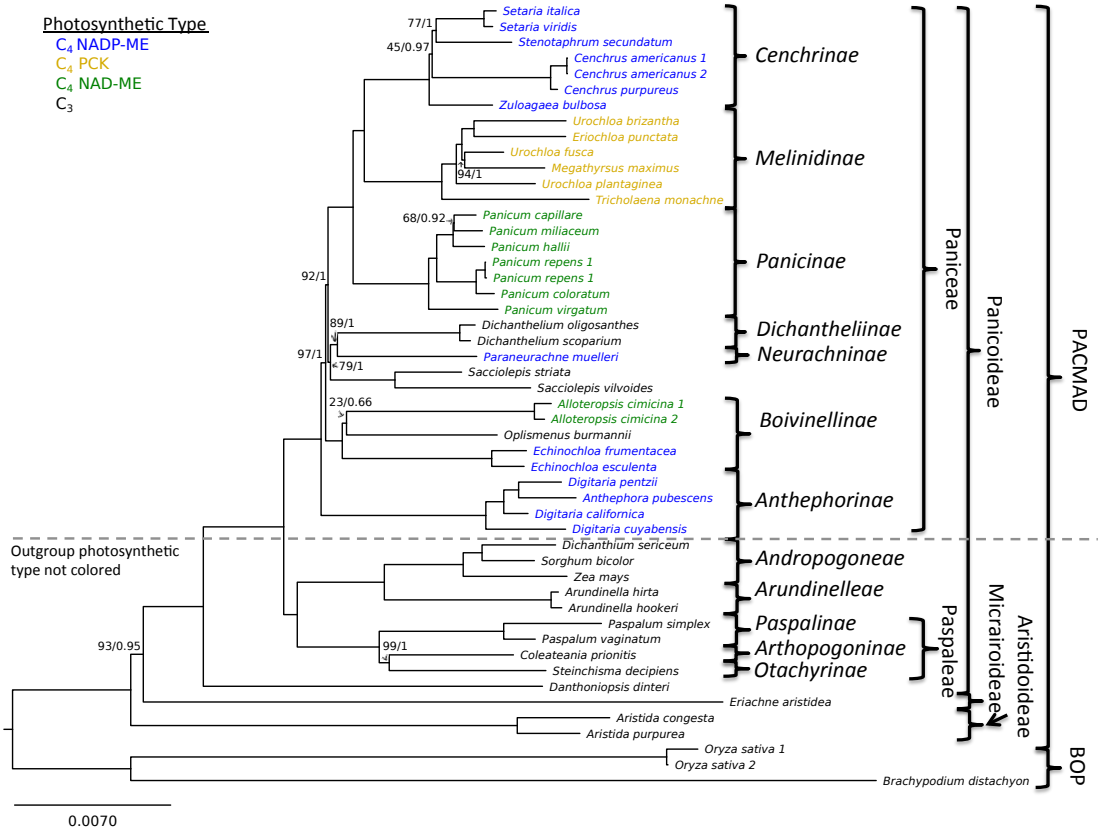


Figure 2.2 - Chloroplast phylogeny of the tribe Paniceae based on 78 loci. Maximum likelihood (ML) tree with both bootstraps (BS) and Bayesian Posterior Probabilities (PP) marked on the branches. Unmarked branches have values of 100 for both BS and PP. Species are colored by photosynthetic sub-type except for outgroups which have not been colored**.** Figure is drawn using the traditional three C_4 sub-type system but can be easily interpreted based on the two sub-type system by replacing PCK with NAD-ME.

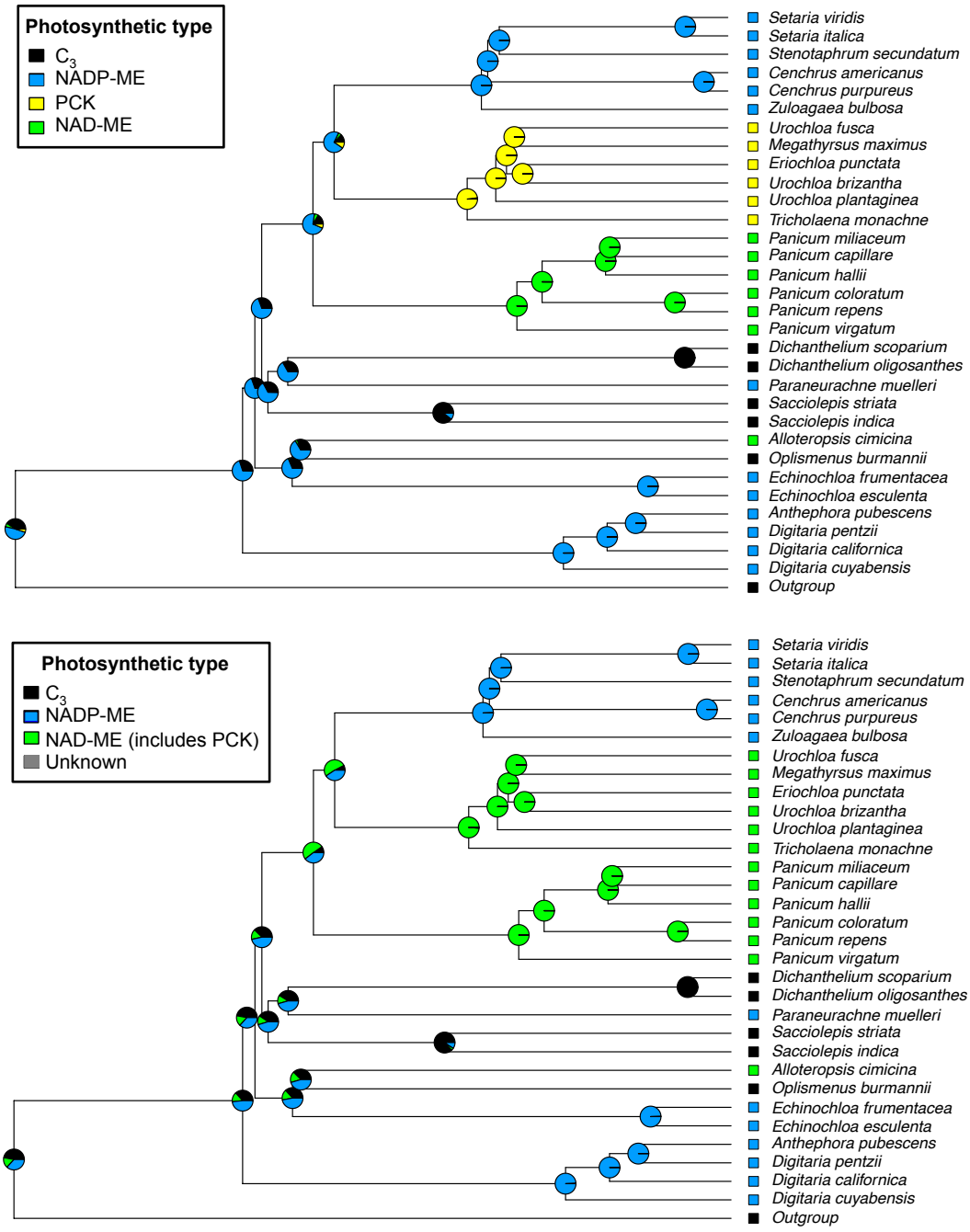


Figure 2.3 - Ancestral state reconstruction of C4 sub-types within the Paniceae. Likelihood based ancestral state reconstructions based on both the classical definition of C_4 photosynthetic sub-types and the two sub-type definition.

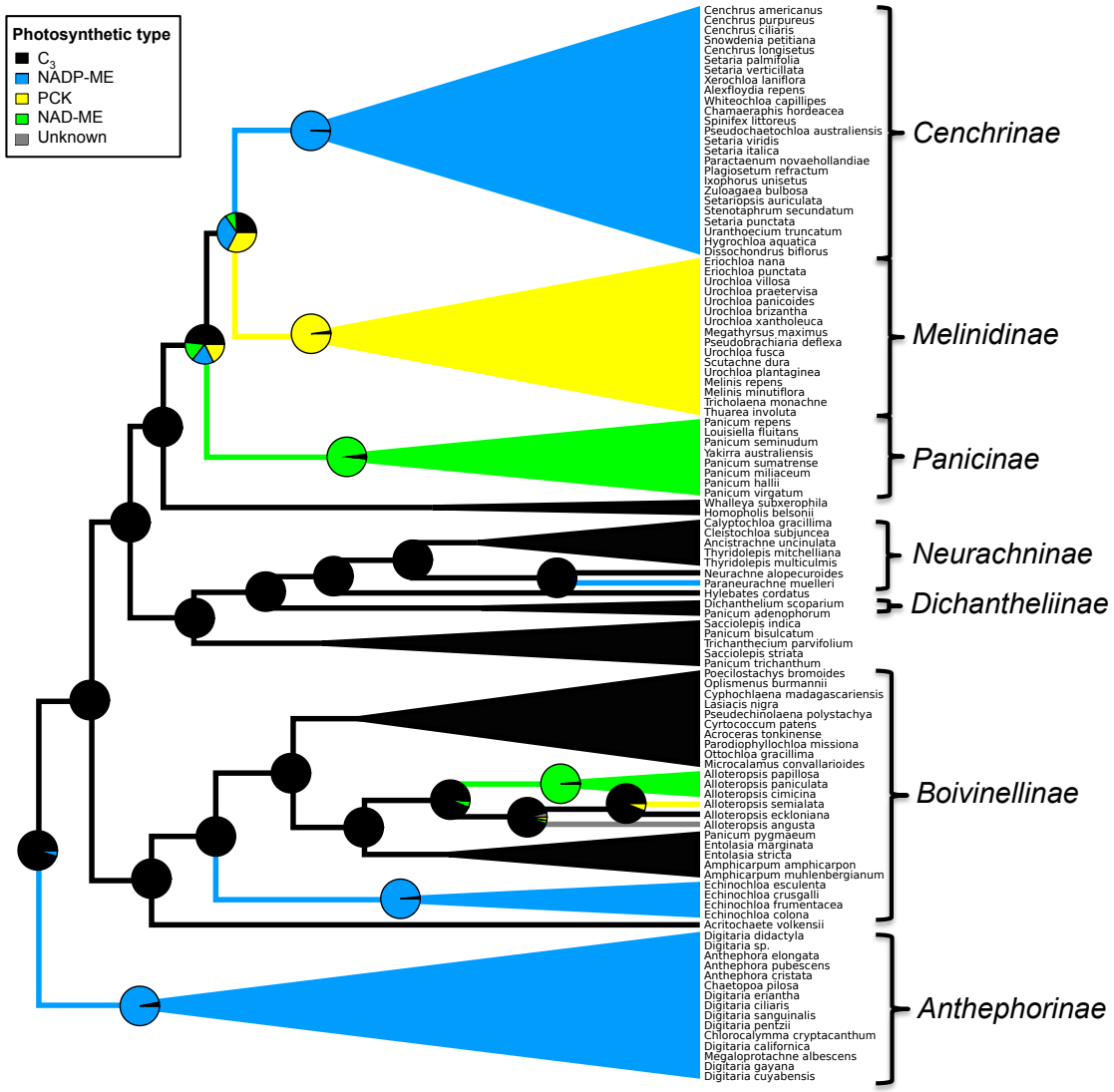


Figure 2.4 - Ancestral state reconstruction of C4 sub-types within the Paniceae (Three C4 sub-types). Likelihood based ancestral state reconstructions based on the classical definition of C_4 photosynthetic sub-types and a combined phylogeny built from both the data generated in this study and that from the GPWG II (2012). Above genus taxonomy labeling has been adjusted to fit recent classification changes (Soreng et al., 2015).

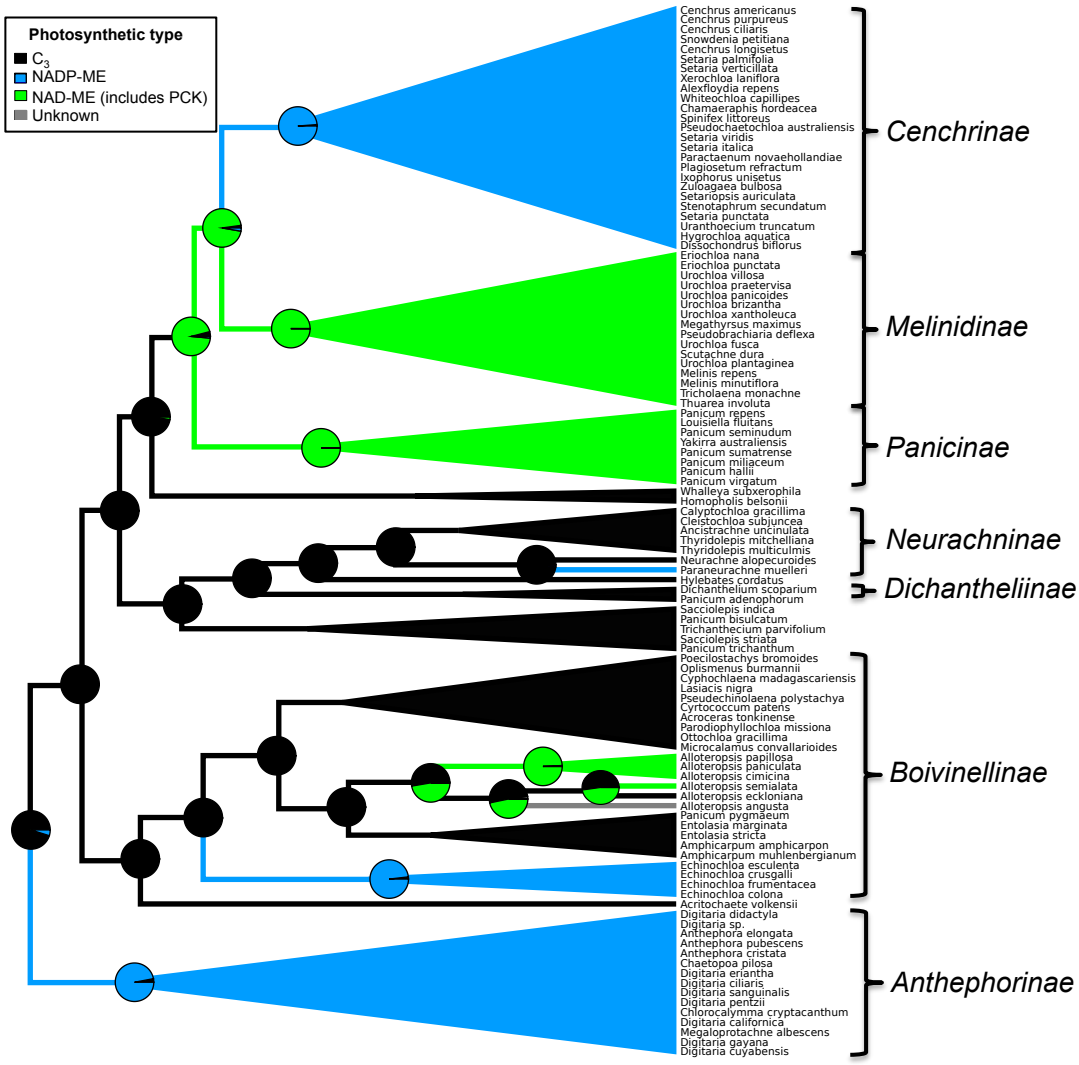


Figure 2.5 - Ancestral state reconstruction of C4 sub-types within the Paniceae (Two C4 sub-types). Likelihood based ancestral state reconstructions based on the two sub-type definition of C_4 photosynthetic sub-types. Mapped onto a combined phylogeny built from both the data generated in this study and that from the GPWG II (2012). Above genus taxonomy labeling has been adjusted to fit recent classification changes (Soreng et al., 2015).

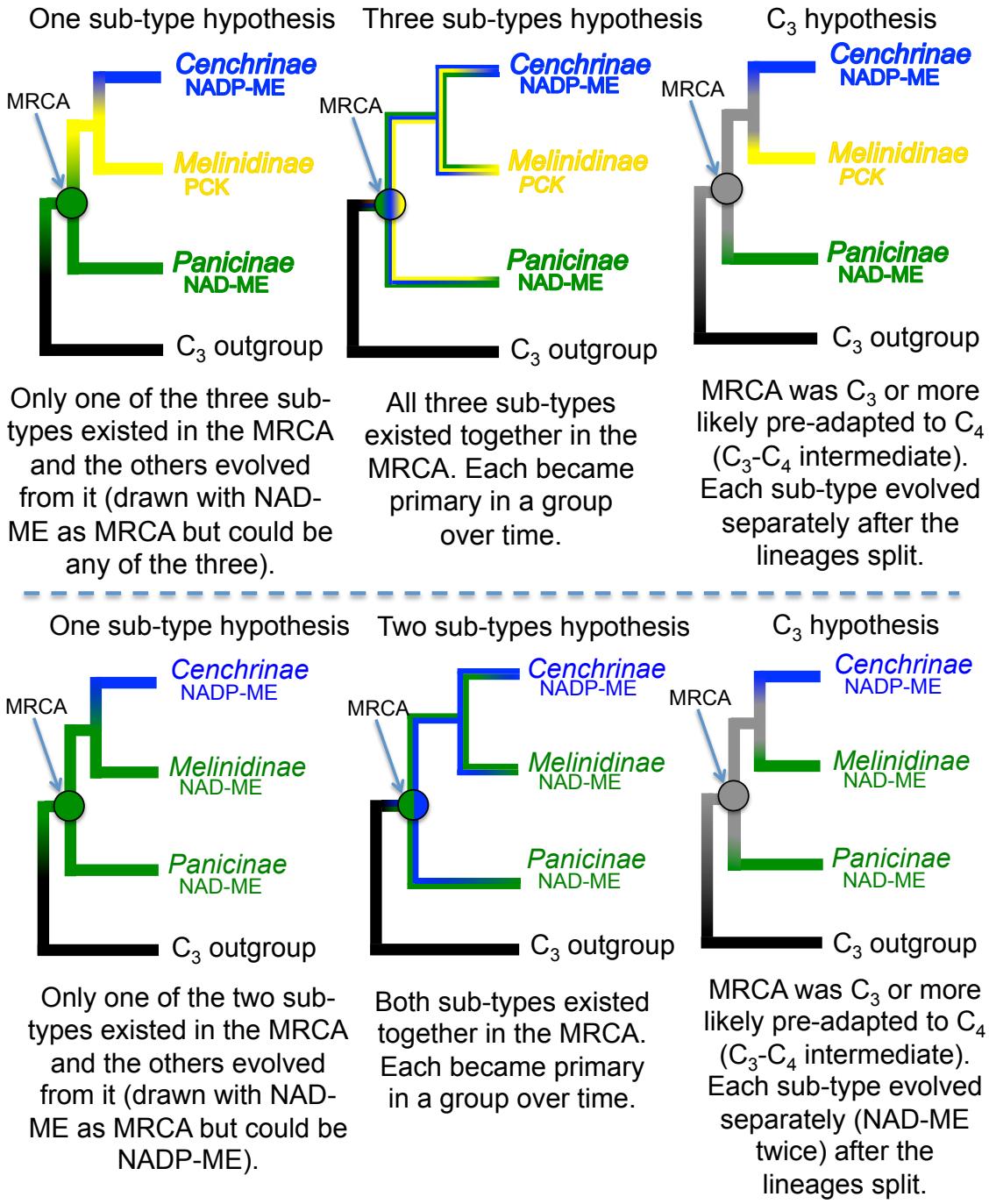
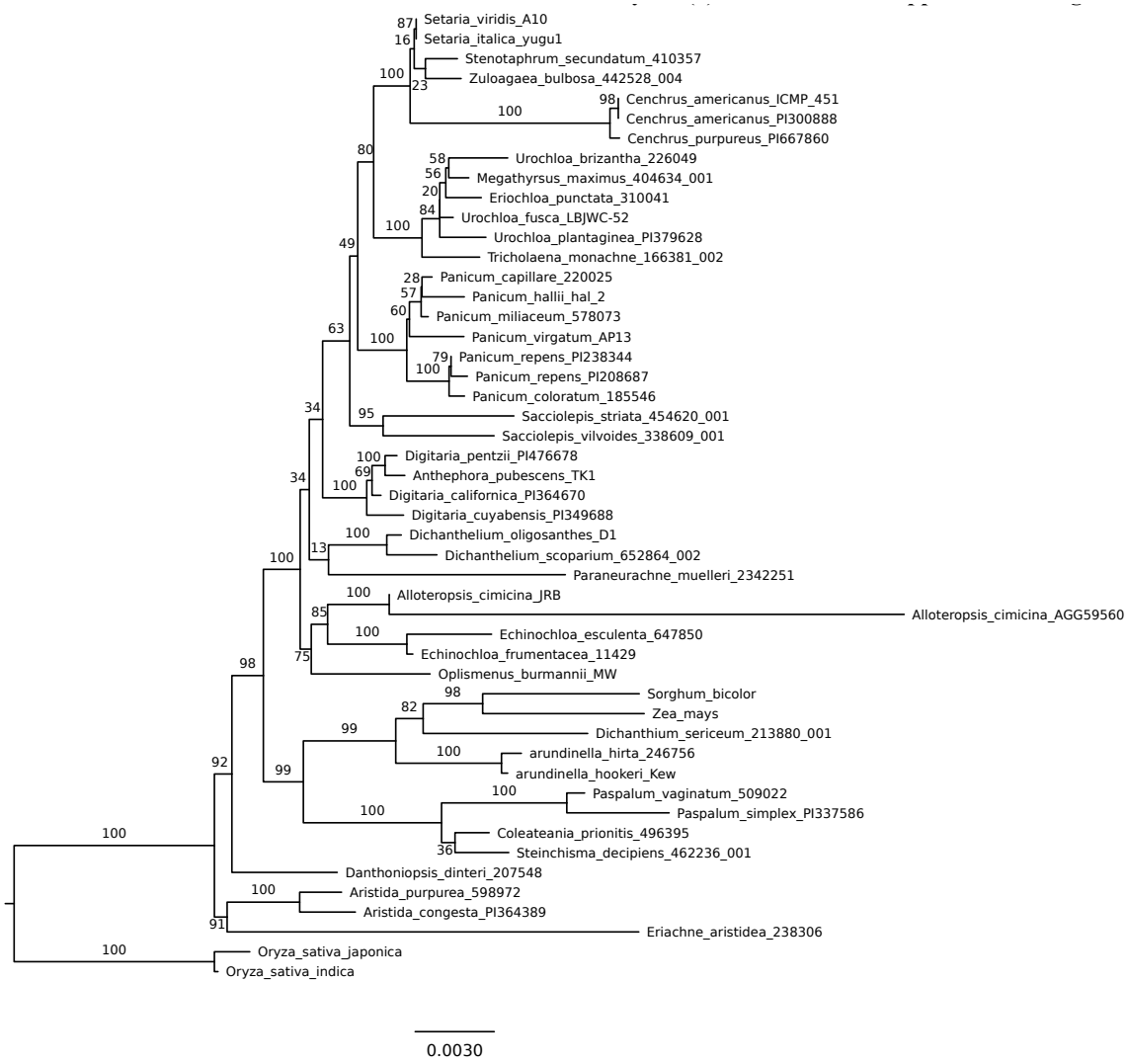


Figure 2.6 - Hypotheses of C4 sub-type evolution within the MPC clade. Generalized hypotheses of how C4 sub-types may have evolved within the MPC clade based on the current chloroplast phylogeny presented in this study. Hypotheses are drawn for both traditional and two sub-type definitions.

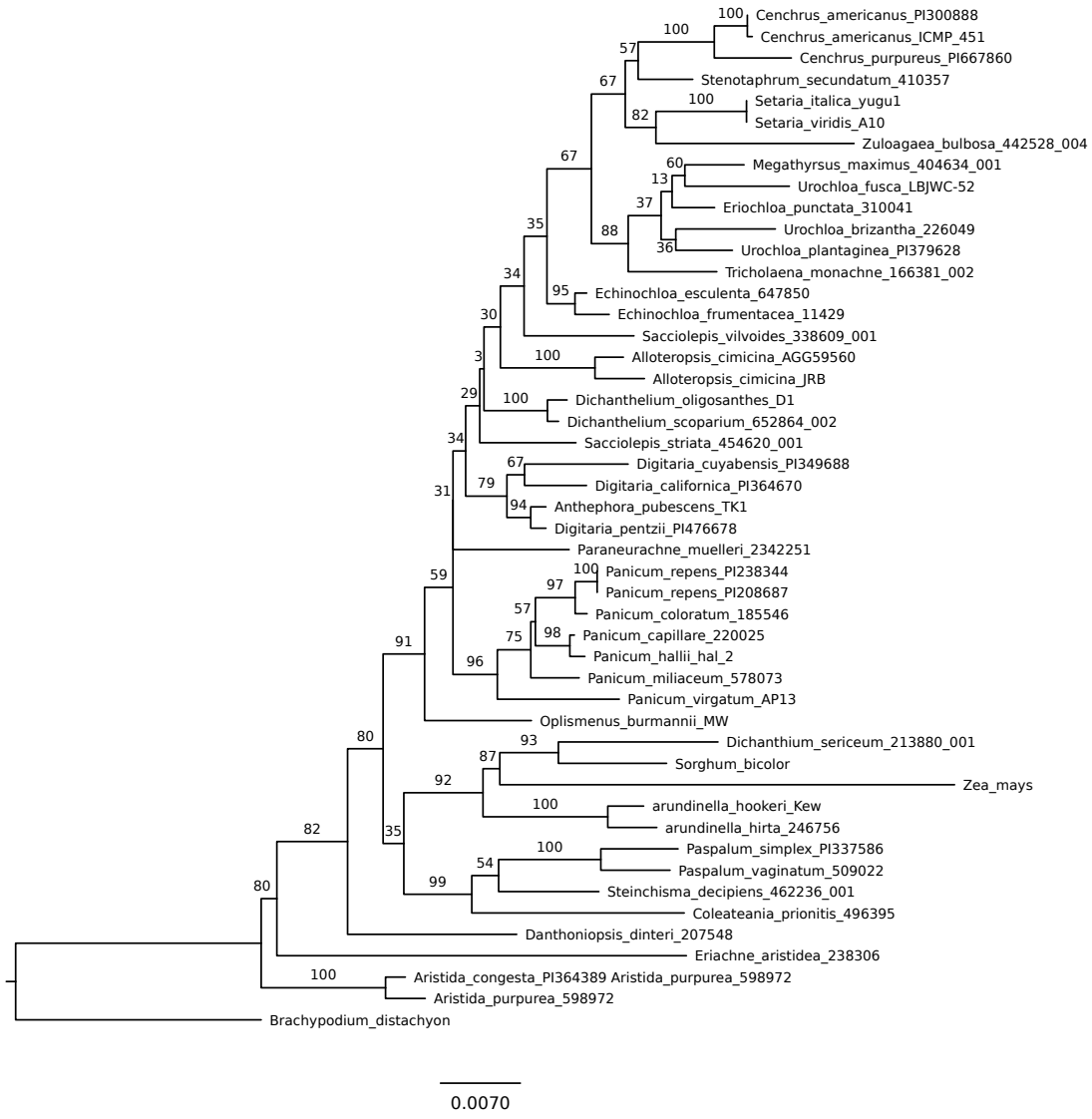
Subtribe	Genus	Species	Authority	Source	ID number	1C Genome size (Mb)	Common name
Cenchrinae	*Cenchrus*	*americanus*	(L.) Morrone	Katrien Devos, UGA	ICMP-451	1,992	pearl millet
Cenchrinae	*Cenchrus*	*americanus*	(L.) Morrone	USDA	PI 300888	2,001	pearl millet
Cenchrinae	*Cenchrus*	*purpureus*	(Schumach.) Morrone	USDA	PI 667860	2,018	elephant grass
Cenchrinae	*Setaria*	*italica*	(L.) P. Beauv.	Thomas Brutnell, Danforth	yugu 1	506	foxtail millet
Cenchrinae	*Setaria*	*viridis*	(L.) P. Beauv.	Thomas Brutnell, Danforth	A10.1	782	green bristlegrass
Cenchrinae	*Stenotaphrum*	*secundatum*	(Walter) Kuntze	USDA	PI 410357	478	St. Augustine grass
Cenchrinae	*Zuloagaea*	*bulbosa*	(Kunth) Bess	USDA	PI 442528	1,877	bulb panic grass
Melinidinae	*Eriochloa*	*punctata*	(L.) Desv. ex Ham.	USDA	PI 310041	1,424	Louisiana cupgrass
Melinidinae	*Megathyrsus*	*maximus*	(Jacq.) B.K. Simon & S.W.L. Jacobs	USDA	PI 404634	1,233	guineagrass
Melinidinae	*Tricholaena*	*monachne*	(Trin.) Stapf & C.E. Hubb.	USDA	PI 166381	1,467	
Melinidinae	*Urochloa*	*brizantha*	(Hochst. ex A. Rich.) R. Webster	USDA	PI 226049	2,591	palisade grass
Melinidinae	*Urochloa*	*fusca*	(Sw.) B.F. Hansen & Wunderlin	USDA	LBJWC-52	400	browntop signalgrass
Melinidinae	*Urochloa*	*plantaginea*	(Link) R.D. Webster	USDA	PI 379628	794	plantain signalgrass
Panicinae	*Panicum*	*capillare*	L.	USDA	PI 220025	459	witchgrass
Panicinae	*Panicum*	*coloratum*	L.	USDA	PI 185546	593	kleingrass
Panicinae	*Panicum*	*hallii*	Vasey	David Lowry, MSU	HAL 2	630	Hall's panicgrass
Panicinae	*Panicum*	*miliaceum*	L.	USDA	PI 578073	1,025	proso millet
Panicinae	*Panicum*	*repens*	L.	USDA	PI 208687*	1,459	torpedo grass
Panicinae	*Panicum*	*repens*	L.	USDA	PI 238344*	1,215	torpedo grass
Panicinae	*Panicum*	*virgatum*	L.	Laura E. Bartley	AP13	1,300	switchgrass
Dichantheliinae	*Dichanthelium*	*oligosanthes*	(Schult.) Gould	Anthony Studer, Danforth	D1	957	Scribner's rosette grass
Dichantheliinae	*Dichanthelium*	*scoparium*	(Lam.) Gould	USDA	PI 652864	896	velvet panicum
Incertae sedis	*Sacciolepis*	*indica*	(L.) Chase	USDA	PI 338609	523	glenwoodgrass
Incertae sedis	*Sacciolepis*	*striata*	(L.) Nash	USDA	NSL 454620	1,582	American cupscale
Neurachninae	*Paraneurachne*	*muelleri*	(Hack.) S.T.Blake	NY	2342251†	unavailable	
Boivinellinae	*Alloteropsis*	*cimicina*	(L.) Stapf	ATCFC	AGG 59560	1,295	summergrass
Boivinellinae	*Alloteropsis*	*cimicina*	(L.) Stapf	James R. Burkhalter, UWF	JRB†	872	summergrass
Boivinellinae	*Echinochloa*	*esculenta*	(A. Braun) H. Scholz	USDA	PI 647850	1,150	Japanese millet
Boivinellinae	*Echinochloa*	*frumentacea*	Link	USDA	Ames 11429	1,590	billion-dollar grass
Boivinellinae	*Oplismenus*	*burmannii*	(Retz.) P. Beauv.	Mark Whitten, FMNH	MW	587	Burmann's basketgrass
Anthephorinae	*Anthephora*	*pubescens*	Nees	Elizabeth Kellogg	TK1	1,623	wool grass
Anthephorinae	*Digitaria*	*californica*	(Benth.) Henrard	USDA	PI 364670*	1,319	Arizona cottontop
Anthephorinae	*Digitaria*	*cuyabensis*	(Trin.) Parodi	USDA	PI 349688*	798	
Anthephorinae	*Digitaria*	*pentzii*	Stent	USDA	PI 476678*	828	slenderstem digitgrass
Outgroups							
Sorghinae	*Dichanthium*	*sericeum*	(R. Br.) A. Camus	USDA	PI 213880	765	silky bluestem
Arundinelleae	*Arundinella*	*hirta*	(Thunb.) Tanaka	USDA	PI 246756	2,614	sae
Arundinelleae	*Arundinella*	*hookeri*	Munro ex Keng	James Schnable	Kew #0050290	unavailable	
Paspalinae	*Paspalum*	*simplex*	Morong	USDA	PI 337586*	1,301	
Paspalinae	*Paspalum*	*vaginatum*	Sw.	USDA	PI 509022	606	seashore paspalum
Otachyriinae	*Steinchisma*	*decipiens*	(Nees ex Trin.) W. V. Br.	USDA	PI 462236	656	
Arthropogoinae	*Coleataenia*	*prionitis*	(Nees) Soreng	USDA	PI 496395	1,656	capim-Santa-Fe
Tristachyideae	*Danthoniopsis*	*dinteri*	(Pilg.) C.E. Hubb.	USDA	PI 207548	760	
	Eriachne	*aristidea*	F. Muell.	USDA	PI 238306	1,255	
	Aristida	*congesta*	Roem. & Schult.	USDA	PI 364389	393	katstertsteekgras
	Aristida	*purpurea*	Nutt.	USDA	PI 598972	3,080	purple threeawn

* Specimins misidentified in USDA collection. Taxon names shown are based on identification from this study.
† DNA Libraries prepared using Nextera kit.

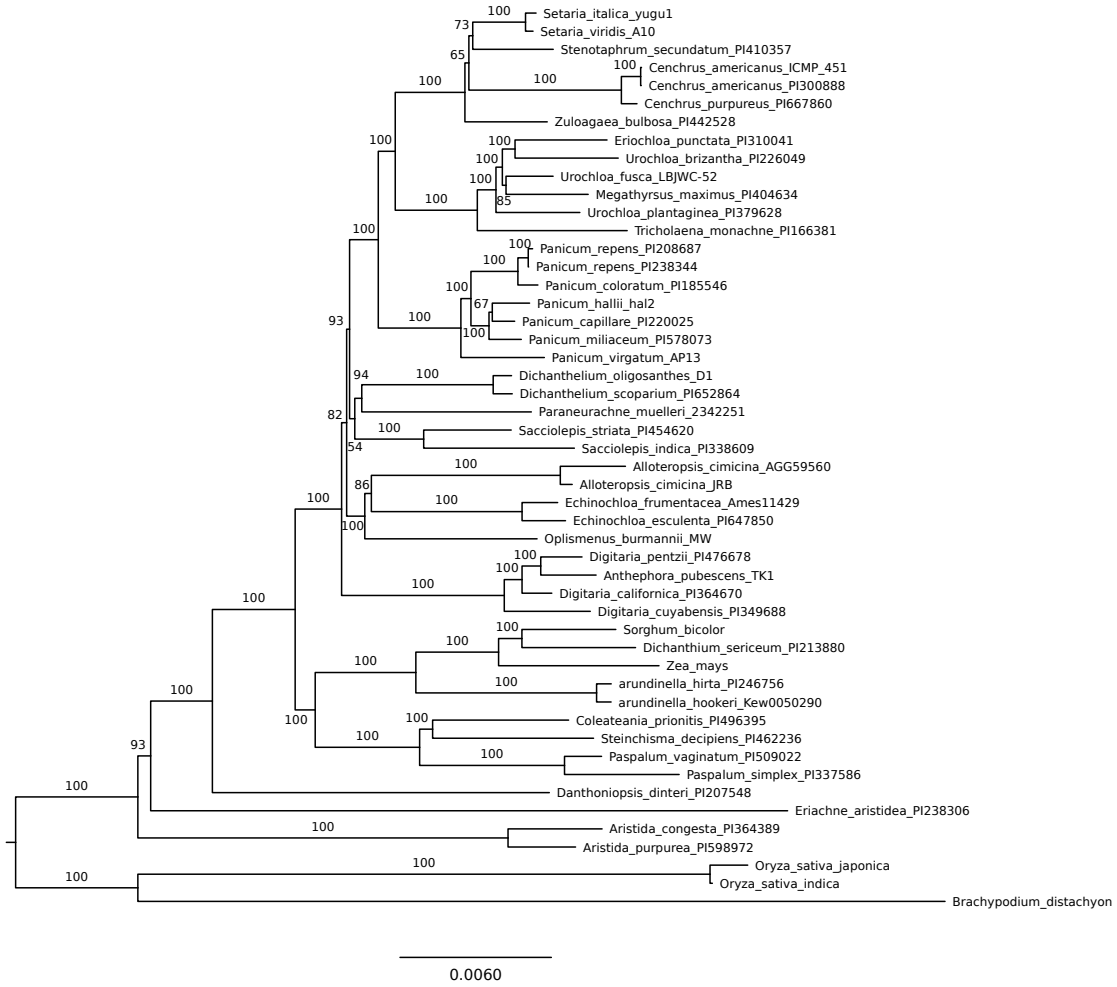
Supplemental Figure S2.1 - Study Material Details. List of all taxa used in this study including their identification numbers, NCBI SRA numbers, common names, source, and estimated genomes sizes.



Supplemental Figure S2.2 - Mitochondrial tree. Maximum likelihood (ML) tree with bootstrap support labels.



Supplemental Figure S2.3 - Nuclear ribosomal tree. Maximum likelihood (ML) tree with bootstrap support labels.



Supplemental Figure S2.4 - Combined (all 102 genes) Paniceae phylogney. Maximum likelihood (ML) tree with bootstrap support labels.

DNA Extraction protocol (Urea method) obtained from Ryan Douglas

Urea extraction buffer (100ml):

42g Urea
7ml 5M NaCl
5ml 1M Tris(Ph=8.0)
4ml 0.5M EDTA
10ml 10% Sarkosyl

1. Grind tissue
2. Place ~500ul of tissue in a 2ml tube.
3. Add 700ul Urea extraction buffer. Vortex. Place on ice.
4. Shake sample at 37°C for 15min.
5. Add 750ul phenol/chloroform/isoamyl alcohol (25:24:1). Vortex for 30 seconds.
6. Shake sample at 37°C for 15min.
7. Centrifuge for 10min at 11,000 rpm.
8. Remove supernatant to a new tube. Discard phenol layer.
9. Add 70ul 3M Sodium Acetate (Ph 5.25) and 700ul isopropyl alcohol. Invert to mix.
10. Centrifuge for 3 min at 13,000 rpm and discard the supernatant. (may need to extend the spin time if pellet doesn't form).
11. Wash twice with 70% EtOH
12. Dry the pellet on a lab bench.
13. Resuspend the pellet in 50ul 1xTE at 4°C.

Supplemental Figure S2.5 - DNA Extraction Protocol (Urea Method).

CHAPTER 3: GENOME-GUIDED PHYLO-TRANSCRIPTOMICS: IMPROVING THE QUALITY OF INFERRED ORTHOLOGS

Washburn, Jacob D.[1]; Schnable, James C.[2,3]; Conant, Gavin C.[4,5]; Brutnell, Thomas P.[3]; Shao, Ying[3,6]; Zhang, Yang[2,3] ; Ludwig, Martha[7]; Davidse, Gerrit[8]; Pires, J. Chris[1]

1) *Division of Biological Sciences, 311 Bond Life Sciences Center, University of Missouri, Columbia, MO, USA 65211*
2) *Agronomy & Horticulture, Beadle Center E207, University of Nebraska-Lincoln, Lincoln, NE, USA 68588-0660*
3) *Donald Danforth Plant Sciences Center, 975 N Warson Rd., St. Louis, MO, USA 63132*
4) Division of Animal Sciences, 920 East Campus Drive, University of Missouri-Columbia, USA
5) Program in Genetics, Bioinformatics Research Center, Department of Biological Sciences, 356 Ricks Hall, North Carolina State University, Raleigh, NC, USA
6) *St. Jude Children´s Research Hospital, MS 342, Room D-4047E, 262 Danny Thomas Place, Memphis, TN, USA 38105*
7) *School of Molecular Sciences, The University of Western Australia (M310), 35 Stirling Highway, Crawley WA 6009, Australia*
8) *Missouri Botanical Garden, P.O. Box 299, St. Louis, Missouri 63166-0299 USA*

Abstract

The past few years have witnessed a paradigm shift in molecular systematics from phylogenetic methods (using one or only a few genes) to those that can be described as phylogenomics (phylogenetic inference with entire genomes). One approach that has recently emerged is phylo-transcriptomics (transcriptome-based phylogenetic inference). As in any phylogenetics experiment, accurate orthology inference is critical to phylo-transcriptomics. To date, most analyses have inferred orthology based either on pure sequence similarity or using gene-tree approaches. The use of conserved genome synteny in orthology detection has been relatively under-employed in phylogenetics, mainly due to the cost of sequencing genomes. While the current trend focuses on the quantity of genes included in an analysis, the use of synteny is likely to improve the quality of ortholog inference. In this study, we combine *de novo* transcriptome data and sequenced genomes from an economically important group of grass species, the tribe Paniceae, to make phylogenomic inferences. This method, which we call "genome-guided phylo-transcriptomics", is compared to other recently published orthology inference pipelines, and benchmarked using a set of sequenced genomes from across the grasses. These comparisons provide a framework for future researchers to evaluate the costs and benefits of adding sequenced genomes to transcriptome data sets. In the case of the grass benchmarking set, twice the percentage of known syntenic orthologs are recovered with the new genome-guided method as are recovered in the other tested methods. The method also enables a new way to investigate and visualize gene tree incongruence; along the length of a chromosome. In addition, this study provides the most comprehensive and

robust nuclear phylogeny of the tribe Paniceae (Poaceae) to date, allowing more informed choices of new genomes to sequence.

Phylogenetic methods have undergone enormous changes over the past few years as the costs of next generation sequencing have declined. Where researchers once spent considerable time designing and testing PCR primers to sequence one or a few genes, it is now becoming common to sequence large numbers of genes, or even whole genomes, for phylogenomic analyses (Cibrián-Jaramillo, et al. 2010, Burleigh, et al. 2011, Lee, et al. 2011, Dunn, et al. 2013a, Salichos and Rokas 2013, Delaux, et al. 2014, Misof, et al. 2014, Yang and Smith 2014, Smith, et al. 2015). In an increasing number of cases, it is possible to build phylogenetic trees based on sequenced genomes, but even these are often re-sequenced or low coverage genomes (Orlando, et al. 2013, Salichos and Rokas 2013, Tsagkogeorga, et al. 2013, Jarvis, et al. 2014, Zhang, et al. 2014, Fontaine, et al. 2015, Foote, et al. 2015, Lamichhaney, et al. 2015, Librado, et al. 2015, Malinsky, et al. 2015, Neafsey, et al. 2015, Lin, et al. 2016). For most groups of eukaryotic organisms, the costs of sequencing and assembling whole genomes remain prohibitive, limiting the applicability of whole genome sequencing for studies that sample large numbers of taxa. Whole genomes are also not generally necessary to allow phylogenomic methods to provide increased resolution of species relationships (Lemmon, et al. 2012). Reduced representation approaches, where part of the genome is excluded from sequencing, allow researchers to obtain sequence data for large numbers of nuclear genes across many species at a relatively low cost and have become increasingly common (Lemmon and Lemmon 2013, Weitemier, et al. 2014, Zimmer and Wen 2015, Budenhagen, et al. 2016, Glenn and Faircloth 2016, McCormack, et al. 2016, Moyle, et al. 2016, Schmickl, et al. 2016).

The current study focuses on improving and testing the constraints of one of these approaches, transcriptome-based phylogenomics. Variations of this method have been applied to a range of organisms and scientific questions (Barker, et al. 2008, Dunn, et al. 2008, Hittinger, et al. 2010, Burleigh, et al. 2011, Wickett, et al. 2011, McKain, et al. 2012, Delaux, et al. 2014, Sveinsson, et al. 2014, Wickett, et al. 2014, Xi, et al. 2014, Cannon, et al. 2015, Edger, et al. 2015, Yang, et al. 2015b, Barker, et al. 2016, Lei and Dong 2016, McKain, et al. 2016, Pease, et al. 2016, Janouškovec, et al. 2017). Transcriptome-based methods differ from other reduced representation approaches in the nature of the gene/transcript ascertainment bias that results. Transcriptomes produce a sampling of transcripts that are biased due to the biology of the organisms under study and the time point(s) and tissue(s) being sampled. Probe-based reduced representation methods on the other hand are biased by the methods used for discovering and choosing the probes. Transcriptome-based approaches to phylogenomics rely on sequencing RNA from multiple taxa at sufficient depth to enable *de novo* assembly of many (usually hundreds to thousands) of transcripts. The resulting transcripts are then used in phylogenetic analyses. The cost of sequencing transcriptomes is, of course, substantially less than that required for whole genomes. Transcriptome-based approaches also require less upfront time investment and a priori knowledge than probe hybridization/sequence capture-based methods. However, they require more bioinformatics time post-sequencing, the reason being that no probe design is required for transcriptome sequencing, but post-sequencing assembly is required. One key limitation of transcriptome-based methods is that they require access to fresh tissue or RNA, and therefore cannot be employed with, for example, museum collections. Conversely, one

advantage of transcriptome-based methods is that the expression data can be used for additional biological analyses beyond phylogenetic inference, as RNA-Seq data is widely used to understand the evolution of gene expression (Dunn, et al. 2013b, Conesa, et al. 2016, Honaas, et al. 2016, Todd, et al. 2016); of course, probe and hybridization-based methods can also be used for other types of functional exploration. Collecting and preserving RNA from fresh or frozen tissue has become routine in many laboratories (Yang, et al. 2017) and, at least in our hands, it is actually easier and less time consuming than DNA sequencing due to streamlined commercial kits (cited in materials and methods below) and the small quantities of RNA required for library preparation.

One area of rapid advancement in transcriptome-based phylogenomics (and most other phylogenomics approaches) is orthology determination. Once transcriptomes are generated and assembled, it is necessary to identify orthologous genes between the various transcriptomes; that is genes that are descended from a single gene copy present in the most recent common ancestor of the species being compared. To date, the most commonly used methods for orthology inference from assembled transcriptomes are based on a multi-step method. First, all-by-all BLAST and the Markov Cluster algorithm (MCL) are used to infer homologous gene sequences (van Dongen 2000, Li, et al. 2003, Camacho, et al. 2009, Duarte, et al. 2010). Second, in some cases, phylogenetic gene trees are built from these homologs and topological features of the trees are used to infer orthologs. Two of the most commonly used platforms for doing this are the Agalma and Yang & Smith pipelines (Howison, et al. 2012, Dunn, et al. 2013a, Yang and Smith 2013, Yang and Smith 2014, Yang, et al. 2015a). These phylogenetically-informed methods have proven effective and become popular in large part because they are computationally

tractable and because they require no *a priori* information about gene order (e.g., they do not require sequenced genomes). One of the major downsides to these methods is the use of an all-by-all BLAST step. Not only can sequence similarity searches be problematic for establishing orthology (Smith and Pease 2016), but when they are performed in a pairwise all-by-all framework they become extremely resource-intensive computationally. Some of these issues can be overcome through the phylogenetically informed approaches described above and the use of parallel computing, but many improvements remain to be made.

An alternative method for orthology inference that has been used in prokaryotes but received relatively little attention in eukaryotic phylogenetics is the use of gene synteny (Bekaert and Conant 2011, Prasanna and Mehra 2013, Wang and Wu 2015). Synteny can be defined as the co-localization of the same gene at similar chromosomal positions across related taxa (Tang, et al. 2008, Bekaert and Conant 2011). Synteny has been compared to a street address system where, if one knows the physical location of a building, it is much easier to find that building than just looking for a building with specific features. Synteny-based orthology determination is then rooted in the assumption that orthologous genes will not only share sequence similarity, but will also reside in similar locations within the genomes of related species (Tang, et al. 2008). Synteny-based methods are widely employed in comparative genomics studies (Cannon and Young 2003, Fu, et al. 2007, Han and Hahn 2009, Jun, et al. 2009, Schnable, et al. 2011, Schnable, et al. 2012a, Schnable, et al. 2012b, Lechner, et al. 2014). They omission of synteny-based approaches in most phylogenetic studies is likely due to the fact that syntenic analysis requires information on gene order in addition to gene sequence, and

information on gene order is not captured by reduced representation methods, including amplification-based, probe-based and transcriptome-based datasets. However, as synteny is widely conserved across many groups of related species (Lyons and Freeling 2008, Lyons, et al. 2008, Tang, et al. 2008), it is possible to use syntenic data from a few genomes as an anchor for reduced representation data, an idea that has not yet been fully explored.

Here we describe the development and implementation of a method we call genome-guided phylo-transcriptomics. This method uses genome-derived syntenic orthologs to anchor transcripts for phylogenetic inference, and is here tested and applied in an economically and scientifically important group of grasses, the tribe Paniceae (Vicentini, et al. 2008, Grass Phylogeny Working Group II 2012, Spriggs, et al. 2014, Washburn, et al. 2015, Burke, et al. 2016, Washburn, et al. 2016). While the method still requires a BLAST step in which transcripts are mapped directly to reference genes that are known to be single-copy orthologs based on synteny, it bypasses the time consuming and error prone all-by-all BLAST and MCL algorithm steps commonly used in current phylo-transcriptomic methods. Furthermore, by removing transcripts that map in multiple copies to the reference ortholog (see Materials and Methods section), one can avoid using BLAST to distinguish between paralogs and orthologs whose sequences are very similar. These are, of course, the sequences for which BLAST is most problematic (Smith and Pease 2016). We hypothesize that the use of a genome-guided method for orthology prediction will result in a greater percentage of "true" orthologs than those predicted by topology-based methods. This decrease in the signal-to-noise ratio in a data set could

have serious impacts given the influence that even a single informative ortholog can have on a phylogenetic analysis (Brown and Thomson 2016).

In addition to the Paniceae data set here generated, we applied the new method to a published dataset from grape (*Vitis vinifera*) and its relatives which covers a wider phylogenetic distance than the tribe Paniceae (Wen, et al. 2013, Yang and Smith 2014). We also constructed and analyzed a data set from several publically available genomes from across the grasses (family Poaceae) and used it to benchmark the method's reliability as compared to orthology inference based entirely on sequenced genomes. The three data sets were analyzed using both this genome-guided method as well as two recently published topology-based approaches for orthology inference with transcriptomes, the Agalma and Yang & Smith pipelines (Howison, et al. 2012, Dunn, et al. 2013a, Yang and Smith 2014).

Materials and Methods

Taxon Sampling and Plant Materials

Forty-five species from across the tribe Paniceae and outgroups were selected for RNA sequencing. Samples were obtained from the sources listed in Supplemental Table S3.1 (available on Dryad), with the majority of samples drawn from the USDA germplasm collection. Most samples were taken from the same plants as those used by Washburn et al. (2015) so results could be directly compared to the chloroplast phylogeny inferred in that study. Plants were grown and sampled in the greenhouse facilities at the University of Missouri, Columbia, MO and the Danforth Center, St.

Louis, MO, with the exception of *Neurachne alopecuroidea* and *Paraneurachne muelleri*, for which RNA samples were obtained from Martha Ludwig, University of Western Australia. Leaf material was sampled from all plants and where possible, shoot, flower, and drought-stressed tissue samples were also taken with the hope of capturing a greater number of unique transcripts. RNA was extracted using the PureLink® RNA Mini Kit (Invitrogen, Carlsbad, CA, USA) or using Roche TriPure (Indianapolis, IN, USA), following the manufacturer's instructions. The grape data set was obtained from NCBI. Details on its generation and record locators can be found in Wen, et al. (2013). The grass genomes and annotation were downloaded from Phytozome (phytozome.jgi.doe.gov) and included *Zea mays* 284 $5b^{+}$ (Schnable, et al. 2009), *Sorghum bicolor* 255 v2.1 (Paterson, et al. 2009), *Setaria italica* 312 v2.2 (Bennetzen, et al. 2012), *Oropetium thomaeum* 386 v1.1 (VanBuren, et al. 2015), *Oryza sativa* 323 v7.0 (Ouyang, et al. 2007), and *Brachypodium distachyon* 283 v2.1 (The International Brachypodium Initiative 2010).

Transcriptome Sequencing

Libraries were prepared using the TruSeq Stranded mRNA Sample Prep Kit (Illumina, Inc., San Diego, CA, USA) or the method described by Wang, et al. (2011). Sequencing was performed at the MU DNA Core facility on the campus of the University of Missouri and at Cornell University's sequencing core facility, and was done on an Illumina Hiseq sequencer with 2 X 100 bp chemistry and six species per lane. Data

generated or used in the study can be found on NCBI SRA under the accession numbers noted in Supplemental Table S3.1.

Sequence Processing

RNA-seq data were quality filtered following standard procedures (Schmieder and Edwards 2011, Babraham Bioinformatics 2015). Transcriptomes were assembled *de novo* using Trinity (Grabherr, et al. 2011, Henschel, et al. 2012, Haas, et al. 2013) and processed as described in Yang and Smith (2013).

The sequenced genomes of *S. bicolor* and *S. italica* were used for syntenic ortholog determination because both are high quality and publically available, they represent an ingroup and outgroup taxa to the tribe Paniceae, and neither genome contains a recent whole genome duplication event (Paterson, et al. 2009, Bennetzen, et al. 2012). Syntenic orthologs between *S. bicolor* and *S. italica* were infered using the SynMap tool in CoGe (https://genomevolution.org/CoGe/) with QuotaAlign set to filter out syntenic paralogous regions using a quota setting of 1:1 (Lyons, et al. 2008, Tang, et al. 2011). Protein sequences of the *S. bicolor* representative orthologs were used as the reference sequence for the remainder of the analyses. The assembled tribe Paniceae transcripts (excluding outgroup transcriptomes) were then mapped to the *S. bicolor* reference orthologs using BLAST with a cutoff E-value of 0.00001 and 85% amino acid identity. When a given *S.bicolor* gene mapped to more than one transcript in a species, all transcripts maping to that gene where discarded. These sequences where then grouped into orthologous sets for each gene and a multiple alignment was created using mafft (Katoh, et al. 2002, Katoh and Standley 2013). In this way, the use of all-by-all BLAST

and the MCL algorithm are completely avoided. After further filtering with phyutility and several scripts from Yang and Smith (2014), concatenated trees, coalescent species trees, and binned coalescent species trees were created using RAxML, ASTRAL, and binning followed by ASTRAL, respectivly (Stamatakis 2006, Mirarab, et al. 2014a, Mirarab, et al. 2014b, Stamatakis 2014) (Figure 3.1). To investigate syntenic block phylogenies, data from the genome-guided gene trees were grouped based on conserved syntenic blocks across the *S. bicolor* and *S. italica* genomes (again obtained from CoGe). Each transcript was mapped to its syntenic block and trees created using RAxML based on concatenated transcripts from each syntenic block. The same method was applied to the grape data set, except that the *V. vinifera* and *Arabidopsis thaliana* genomes where used and the E-value and protien identity cutoffs where lowered to 0.0001 and 75%, respectively, to account for the increased phylogenetic distances represented in the grape data set. Scripts and instructions for the genome-guided method are available at: bitbucket.org/washjake/transcriptome_phylogeny_tools.

Two gene tree topology-based approaches to orthology inference where also used for comparison: the Agalma pipeline (version 0.5.0) by Dunn, et al. (2013a) and the Ya Yang pipeline (Yang and Smith 2013, Yang and Smith 2014). As above, RAxML, ASTRAL, and binning combined with ASTRAL were used to infer phylogenies.

For the grass data used for benchmarking, several additional analyses were run. Single copy syntenic orthologs were found in a pairwise fashion between *O. sativa* and each of the other genomes using CoGe as described above. These orthologs where used to create a set of high-confidence, fully synteny-based, one-to-one orthologs across the grasses. While this set does not include all possible single-copy orthologs, it does include

all of them for which we can have high confidence based on the available data and current methods, and represents the closest thing to a gold standard ortholog set currently possible for the grasses, with rice as the reference. We refer to this as the benchmarking data set.

Each of the ortholog inference methods described above was then run using the transcriptomes generated by the genome sequencing projects referenced above. In this way, the transcripts could be followed by name through the pipelines (except for the Agalma method for which this could not be easily accomplished due to the way the pipeline is packaged). Ortholog sets derived from the genome-guided method and the Yang and Smith method were then compared to the benchmarking set to determine how many orthologs each method was able to find in common with the benchmark orthologs.

Results

For species tree inference in the tribe Paniceae, the genome-guided method provided similar numbers of orthologous genes to both the Agalma and Yang & Smith methods at a 90% matrix occupancy cut-off (Table 3.1). However, for the full matrix runs, when any orthologous gene without all species represented was discarded, the genome-guided method returned fewer orthologs than the other two methods. This is probably due to the genome-guided method not using transcripts that map to the same ortholog. The genome-guided method however, produced more consistent tree topologies than the topology-based methods. For example, all species trees (concatenated, coalescent, binned, and with multiple matrix occupancies and taxonomic inclusion) built with the genome-guided orthology pipeline agreed in their subtribe level topologies. The

topology-based methods on the other hand, occasionally produced conflicting subtribe-level topologies. In other words, the topology-based methods were more sensitive to perturbations in taxonomic inclusion than the genome-guided method. The genome-guided method was also many times faster than the topology-based methods (Table 3.2).

Figure 3.2 shows what we consider to be the most conservative and best estimate of the Paniceae nuclear species tree, based on currently available data. This tree places Anthephorinae as direct sister to the MPC clade (subtribes Melinidinae, Panicinae, and Cenchrinae), which, although different from published chloroplast trees (Grass Phylogeny Working Group II 2012, Washburn, et al. 2015, Burke, et al. 2016), is consistent with the combined nuclear-chloroplast topology reported by Vicentini, et al. (2008).

As mentioned, the topology-based approaches (Yang & Smith and Agalma) generally resulted in the same tree topology as the genome-guided method (Figure 3.3a). However, in some cases, depending on the taxon sampling included in the analysis, an alternative topology was obtained from these methods. This topology placed the subtribe Anthephorinae together with the Neurachninae and *Sacciolipis* lineages as sister to the MPC clade (Figure 3.3b). Internode certainty (IC) scores for the main conflicting node in both the primary and secondary topologies were close to zero and in some cases even negative, suggesting high levels of gene tree incongruence (Salichos and Rokas 2013, Salichos, et al. 2014, Kobert, et al. 2016). Both the genome-guided method, and the topology-based methods included genes representative of each of the *Sorghum bicolor* chromosomes and the major *Setaria italica* scaffolds, indicating that the sampled genes from both methods came from across the entire genome (Supplemental Table S3.2).

To further dissect the causes of gene tree incongruence within the Paniceae, the tree binning scripts described by Mirarab, et al. (2014a) where used to separate groups of genes with distinct evolutionary histories. This method allows one to set a significance threshold at which branches can be considered high confidence, and then compare large sets of gene trees for compatibility with each other. Different cut-off values were tested for this analysis, almost always (see exception below when a cut-off of 100 was used) resulting in several hundred distinct tree topologies that were incompatible with each other.

When a bootstrap cutoff value of 100 was used as the threshold, indicating that only gene tree branches with 100 percent bootstrap support were considered, the synteny-based data set still placed the trees into 18 unique topology groups. These eighteen topologies were then compared visually and examined for differences that could directly affect the relationship between the MPC clade and the subtribe Anthephorine. Of the eighteen (2211 total genes) topologies, eight topologies (981 total genes) showed strong support for the inclusion of Anthephorine within the MPC (as in the primary topology described above and shown in Figures 3.2 and 3.3b), five topologies (615 total genes) showed strong support for Anthephorine as sister to Neurachninae and *Sacciolipis* (as in the secondary topology described above and shown in Figure 3.3b), and one topology (123 total genes) agreed with the chloroplast phylogeny from Washburn, et al. (2015) (Figure 3.3c). The remaining four topologies (492 total genes) had low support for this area of the tree.

Another approach we developed to dissect gene tree incongruence consisted of building trees based on the combination of genes that share a similar physical location. A

recent study was able to find likely introgression events using a non-overlapping window approach and constructing trees based on 1 Mb and 100 kb blocks of genes (Pease, et al. 2016). Because of the genome-guided approach, we were able to group genes into more biologically relevant blocks, namely blocks that are syntenically conserved between ingroup and outgroup taxa. The appearance of a block of genes sharing the same phylogeny, which differs from the species phylogeny, might suggest hybridization/introgression within a group as recently diverged as the Paniceae, but ILS could also produce these types of blocks.

Many syntenic block phylogenies were inconclusive in that they yielded topologies that had little similarity to any of the previously described or published species trees. This seemed to be correlated with the number of genes in a syntenic block in that blocks with more genes generally (but not always) provided a resolved phylogeny that was similar in the placement of the subtribes Melinidinae, Panicinae, Cenchrinae, Anthephorinae, Neurachninae, and the *Sacciolipis* lineage to one of the three phylogenies in Figure 3.3. When these blocks and their topologies were mapped to an ideogram of the *S. italica* chromosomes a striking patchwork of differing syntenic block histories was revealed (Figure 3.3d). To further investigate wither or not syntenic blocks have distinct tree topologies we used Robinson-Foulds (RF) distances as implemented in the ETE Toolkit (Huerta-Cepas, et al. 2016). By computing pairwise RF distances for all genes in a given block we created a tree distribution for each of the blocks. We then took the complete set of gene trees (those from all blocks) and randomly re-assigned them to blocks eighty thousand times, each time computing the pairwise RF distance for each block. In this way, we created a simulated "random distribution" of trees for each block

that could be used as the null distribution in a statistical test comparing the observed pair-wise RF distances in a block to the simulated distribution under the null hypothesis that all blocks share the same tree distribution. Of the 79 blocks, 15 had distributions that were significantly different then their respective simulated distributions at a significance level of *alpha* < 0.001 (Figure 3.3d). Hence, it appears that these blocks have a distribution of tree distances smaller than that expected. This observation implies that at least some local regions of the genome have similar evolutionary histories relative to the genome as a whole, either because of locally-coherent ILS or hybridization.

To further benchmark the Genome-Guided method here developed we applied it to two additional data sets. We used publically available sequenced grass genomes to compare our method with the Agalma and Yang and Smith methods. Syntenic relationships between the genomes were used to construct a list of high confidence single copy orthologs across the grasses. This list then served as a benchmark to which the orthologs predicted by each of the methods could be compared. Of course, this list does not contain all orthologs across the grasses, so it cannot tell us anything about the validity of ortholog combinations predicted for genes not found in the list. However, it can tell us when each ortholog prediction method correctly places orthologs in its list, and when it incorrectly identifies paralogs as orthologs. Therefore, we think that these comparisons are informative as to the reliability of different orthology assignment methods.

All three orthology detection pipelines and the synteny-derived benchmarking set generated the same tree topology, in agreement with previous phylogenetic studies, with high confidence (Grass Phylogeny Working Group II 2012). Gene by gene comparisons between each method and the benchmarking set show that the genome-guided method

recovers a much higher percentage of ortholog gene trees that agree, in terms of which genes are included, with the benchmarking set than either the Yang & Smith 1 to 1 or MO Methods (62%, 32%, and 31% respectively with a species cutoff of four and genes not found in the benchmarking set excluded. See Table 3.3). Because of the way Agalma is packaged, we were unable to modify its code to include it in this comparison, but it would likely perform similarly to the Yang & Smith method as it uses similar approaches. Beyond the orthologous genes in the benchmarking set, the Yang & Smith methods also include as many as 2,116 additional ortholog gene trees. These trees are based on genes which our direct synteny comparisons did not find. They may or may not be based on correct orthology assignment, but because they are not in the benchmarking set they could not be evaluated here.

We also analyzed previously published data from grape and its relatives in order to benchmark the method and explore the phylogenetic distance it is capable of spanning. The grape data set performed similarly to the Paniceae dataset we generated in terms of the amount of time it took to perform the genome-guided method versus the other methods. The number of orthologs retrieved was much smaller for the genome-guided method than it was for the two topology-based methods (Supplemental Table S3.3). This dearth of orthologs is likely due to the simple fact that syntenic relationships are expected to break down as the evolutionary distance between two species increases. Grape and Arabidopsis likely diverged between 69-150 million years ago (m.y.) with most estimates around 100 m.y., while Sorghum and Seteria probably diverged between 25-40 m.y. (Vicentini, et al. 2008, Grass Phylogeny Working Group II 2012, Stevens 2017). Even with substantially fewer genes, the genome-guided method still predicted similar

topologies to those of the other two methods and those previously published (Figure 3.4) (Yang and Smith 2014).

Discussion

Genome-Guided versus Topology-Based Methods

Phylogenetic consistency, broadly defined as convergence on the "correct" tree topology with increasing data, is a well-established phylogenetic accuracy assessment criterion (Hillis 1995, Huelsenbeck 1995, Nabhan and Sarkar 2011). The Genome-guided method proposed here consistently inferred the same subspecies level tree topology regardless of the matrix occupancy used, the tree building approach applied, and the number of taxa included. The topology-based approaches also resulted in the same tree topology in most cases, however, when the number of taxa were reduced to 33 by removing species near the main areas of conflict, the topology-based approaches no longer produced consistent results while the Genome-guided method continued to produce the same topology. In general, the inclusion of more taxa, which better represent the diversity of a group of organisms, will increase the accuracy of phylogenetic inference (Hillis, et al. 2003, Havird and Miyamoto 2010, Nabhan and Sarkar 2011). It then follows that the topology found by both genome-guided and topology-based methods, when all taxa were included, is likely to be the topological estimate nearest to the true species history. This implies that the genome-guided method should be able to infer that topology with less data than the topology-based methods require for similar confidence and accuracy.

Additionally, the grass benchmarking data set comparisons indicate that, of the orthologs we know with high confidence, the genome-guided method predicts a higher percentage of them correctly than does the Yang & Smith method.

Computational Times and Resources

Orthology inference with the genome-guided pipeline is also many times faster than the topology-based methods and, except for the CoGe step, can be run efficiently on a standard desktop computer; something not possible with either of topology-based methods. This computational efficiency results from the fact that the genome-guided method does not require all-by-all BLAST or iterative tree pruning steps. The CoGe step is also very simple and straight-forward to run, as is the process of loading new genomes into the CoGe database. CoGe also has the capacity for uploading and analyzing private genomes without making them public and is exceptionally well documented.

A natural downside of the genome-guided method is the need for two genomes that span the taxonomic clade one is working with. While this approach could be used with only one genome or even a list of genes from a *de novo* transcriptome assembly, doing so negates its benefits and will increase the likelihood of including paralogs in the analysis. In these situations, topology-based methods are probably the best analysis choice.

Gene Tree Incongruence

Based on both the binning analysis and the syntenic block trees, we conclude that the secondary topology, or at least the differential placement of the Anthephorine relative

to the MPC, is not an artifact of the topology-based methods, but is supported by an appreciable number of genes regardless of the orthology determination method employed. The different topologies of these genes may result from either ILS or post-speciation hybridization, or both.

The small numbers of transcripts representing many of the syntenic blocks in Figure 3.3, likely contributed to an inability to infer well supported phylogenies for some of the blocks. However, RF based topology distribution tests confirmed that tree topology distributions in at least certain areas of the genome are likely more similar for genes in a syntenic block than they are across the whole genome. This type of local-synteny analysis should become even more informative in future studies as more sequenced genomes are generated and included in phylogenetic inference. These types of analyses are also not limited to transcriptomic data but have the potential to add value to other data sets generated with probe/hybridization based data collection methods, as long as one or more sequenced genomes exist within the taxonomic group being studied.

Phylogeny of the Tribe Paniceae

The nuclear phylogeny of the Tribe Paniceae produced in this study is consistent with that produced in a previous study. However, that study was only able to sample one nuclear gene and because the inferred topology was incongruent with the many chloroplast phylogenies of the group, it was generally dismissed. This study demonstrates that in fact the nuclear phylogeny of the Paniceae is very different than the chloroplast one, and that those differences are not due to signals in one or a few genes, but are wide

spread across the genome. This study also shows that while that original topology, based on only one nuclear gene is supported by many other genes, not all nuclear genes agree with it, and in fact a significant minority of the genes are incongruent with that topology.

The differences between the nuclear and chloroplast phylogenies shown here are critical to both basic and applied questions within the tribe Paniceae. For example, investigations within the tribe of the evolution of C_4 photosynthesis, a trait with great economic importance, have focused on the MPC clade at the exclusion of the subtribe Anthephorinae (Grass Phylogeny Working Group II 2012, Washburn, et al. 2015). Choices about resource investment, such as which genomes to sequence, have also been based almost exclusively on the chloroplast phylogeny (Studer, et al. 2016). Given our results, further resource investment in Paniceae (at least for the purpose of studying C_4 photosynthesis) should be directed within the genus *Sacciolepis* or a close relative to it and the subtribe Anthephorinae. We suggest *Sacciolepis indica* as a model C_3 species for further study as it is a close relative the MPCA clade in both chloroplast and nuclear phylogenies, has a genome size of approximately 523Mb, and is easily self-pollinated (Washburn, et al. 2015). An ideal Anthephorinae species for further investment is less clear, but *Digitaria cuyabensis* has an approximate genome size of 798Mb making it a good candidate for genome sequencing (Washburn, et al. 2015). Species within the Crabgrass complex, which includes several different species in the genus *Digitaria*, might also be good candidates for resource investment due to their economic importance as a noxious weed. Genome-guided phylo-transcriptomics allows for a more informed way to choose future genomes to sequence because, as is shown here, the nuclear-gene trees may differ from the organellar gene trees.

Funding

This work was supported by the National Science Foundation (DEB Award no. 1501406), the University of Missouri Research Board, the University of Missouri Mizzou Advantage, and the Sigma Xi Grants-in-Aid of Research Program.

Acknowledgments

The authors thank Elizabeth Kellogg for species sampling advice, Antonis Rokas, Casey Dunn, Stephen A. Smith, Ya Yang, and James Pease for review and/or comments that greatly enhanced the manuscript, and xxx reviewers for their thoughtful comments.

References

Babraham Bioinformatics. 2015. FastQC A quality control tool for high throughput sequence data.

Barker MS, Kane NC, Matvienko M, Kozik A, Michelmore RW, Knapp SJ, Rieseberg LH. 2008. Multiple Paleopolyploidizations during the Evolution of the Compositae Reveal Parallel Patterns of Duplicate Gene Retention after Millions of Years. Mol. Biol. Evol. 25:2445-2455.

Barker MS, Li Z, Kidder TI, Reardon CR, Lai Z, Oliveira LO, Scascitelli M, Rieseberg LH. 2016. Most Compositae (Asteraceae) are descendants of a paleohexaploid and all share a paleotetraploid ancestor with the Calyceraceae. Am. J. Bot. 103:1203-1211.

Bekaert M, Conant GC. 2011. Copy Number Alterations among Mammalian Enzymes Cluster in the Metabolic Network. Mol. Biol. Evol. 28:1111-1121.

Bennetzen JL, Schmutz J, Wang H, Percifield R, Hawkins J, Pontaroli AC, Estep M, Feng L, Vaughn JN, Grimwood J, Jenkins J, Barry K, Lindquist E, Hellsten U, Deshpande S, Wang X, Wu X, Mitros T, Triplett J, Yang X, Ye CY, Mauro-Herrera M, Wang L, Li P, Sharma M, Sharma R, Ronald PC, Panaud O, Kellogg EA, Brutnell TP, Doust AN, Tuskan GA, Rokhsar D, Devos KM. 2012. Reference genome sequence of the model plant Setaria. Nat. Biotechnol. 30:555-561.

Brown JM, Thomson RC. 2016. Bayes factors unmask highly variable information content, bias, and extreme influence in phylogenomic analyses. Syst. Biol.:syw101.

Budenhagen C, Lemmon AR, Lemmon EM, Bruhl J, Cappa J, Clement WL, Donoghue M, Edwards EJ, Hipp AL, Kortyna M, Mitchell N, Moore A, Prychid CJ, Segovia-Salcedo MC, Simmons MP, Soltis PS, Wanke S, Mast A. 2016. Anchored Phylogenomics of Angiosperms I: Assessing the Robustness of Phylogenetic Estimates. bioRxiv.

Burke SV, Wysocki WP, Zuloaga FO, Craine JM, Pires JC, Edger PP, Mayfield-Jones D, Clark LG, Kelchner SA, Duvall MR. 2016. Evolutionary relationships in Panicoid grasses based on plastome phylogenomics (Panicoideae; Poaceae). BMC Plant Biol. 16:1-11.

Burleigh JG, Bansal MS, Eulenstein O, Hartmann S, Wehe A, Vision TJ. 2011. Genome-Scale Phylogenetics: Inferring the Plant Tree of Life from 18,896 Gene Trees. Syst. Biol. 60:117-125.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden T. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10:421.

Cannon SB, McKain MR, Harkess A, Nelson MN, Dash S, Deyholos MK, Peng Y, Joyce B, Stewart CN, Rolf M, Kutchan T, Tan X, Chen C, Zhang Y, Carpenter E, Wong GK-S, Doyle JJ, Leebens-Mack J. 2015. Multiple Polyploidy Events in the Early Radiation of Nodulating and Nonnodulating Legumes. Mol. Biol. Evol. 32:193-210.

Cannon SB, Young ND. 2003. OrthoParaMap: Distinguishing orthologs from paralogs by integrating comparative genome data and gene phylogenies. BMC Bioinformatics 4:35.

Cibrián-Jaramillo A, De la Torre-Bárcena JE, Lee EK, Katari MS, Little DP, Stevenson DW, Martienssen R, Coruzzi GM, DeSalle R. 2010. Using Phylogenomic Patterns and Gene Ontology to Identify Proteins of Importance in Plant Evolution. GBE 2:225-239.

Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Szcześniak MW, Gaffney DJ, Elo LL, Zhang X, Mortazavi A. 2016. A survey of best practices for RNA-seq data analysis. Genome Biol. 17:1-19.

Delaux P-M, Varala K, Edger PP, Coruzzi GM, Pires JC, Ané J-M. 2014. Comparative Phylogenomics Uncovers the Impact of Symbiotic Associations on Host Genome Evolution. PLoS Genet. 10:e1004487.

Duarte JM, Wall PK, Edger PP, Landherr LL, Ma H, Pires JC, Leebens-Mack J, dePamphilis CW. 2010. Identification of shared single copy nuclear genes in Arabidopsis, Populus, Vitis and Oryza and their phylogenetic utility across various taxonomic levels. BMC Evol. Biol. 10:1-18.

Dunn C, Howison M, Zapata F. 2013a. Agalma: an automated phylogenomics workflow. BMC Bioinformatics 14:330.

Dunn CW, Hejnol A, Matus DQ, Pang K, Browne WE, Smith SA, Seaver E, Rouse GW, Obst M, Edgecombe GD, Sorensen MV, Haddock SHD, Schmidt-Rhaesa A, Okusu A, Kristensen RM, Wheeler WC, Martindale MQ, Giribet G. 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. Nature 452:745-749.

Dunn CW, Luo X, Wu Z. 2013b. Phylogenetic Analysis of Gene Expression. Integr. Comp. Biol. 53:847-856.

Edger PP, Heidel-Fischer HM, Bekaert M, Rota J, Glöckner G, Platts AE, Heckel DG, Der JP, Wafula EK, Tang M, Hofberger JA, Smithson A, Hall JC, Blanchette M, Bureau TE, Wright SI, dePamphilis CW, Eric Schranz M, Barker MS, Conant GC, Wahlberg N, Vogel H, Pires JC, Wheat CW. 2015. The butterfly plant arms-race escalated by gene and genome duplications. Proc. Natl. Acad. Sci. USA 112:8362-8366.

Fontaine MC, Pease JB, Steele A, Waterhouse RM, Neafsey DE, Sharakhov IV, Jiang X, Hall AB, Catteruccia F, Kakani E, Mitchell SN, Wu Y-C, Smith HA, Love RR, Lawniczak MK, Slotman MA, Emrich SJ, Hahn MW, Besansky NJ. 2015. Extensive introgression in a malaria vector species complex revealed by phylogenomics. Science 347.

Foote AD, Liu Y, Thomas GWC, Vinar T, Alfoldi J, Deng J, Dugan S, van Elk CE, Hunter ME, Joshi V, Khan Z, Kovar C, Lee SL, Lindblad-Toh K, Mancia A, Nielsen R, Qin X, Qu J, Raney BJ, Vijay N, Wolf JBW, Hahn MW, Muzny DM, Worley KC, Gilbert MTP, Gibbs RA. 2015. Convergent evolution of the genomes of marine mammals. Nat. Genet. 47:272-275.

Fu Z, Chen X, Vacic V, Nan P, Zhong Y, Jiang T. 2007. MSOAR: A High-Throughput Ortholog Assignment System Based on Genome Rearrangement. J. Comput. Biol. 14:1160-1175.

Glenn TC, Faircloth BC. 2016. Capturing Darwin's dream. Molecular Ecology Resources 16:1051-1058.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 29:644-652.

Grass Phylogeny Working Group II. 2012. New grass phylogeny resolves deep evolutionary relationships and discovers C4 origins. New Phytol. 193:304-312.

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, MacManes MD, Ott M, Orvis J, Pochet N, Strozzi F, Weeks N, Westerman R, William T, Dewey CN, Henschel R, LeDuc RD, Friedman N, Regev A. 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat. Protocols 8:1494-1512.

Han MV, Hahn MW. 2009. Identifying Parent-Daughter Relationships Among Duplicated Genes. Pacific Symposium on Biocomputing 14:114-115.

Havird JC, Miyamoto MM. 2010. The importance of taxon sampling in genomic studies: An example from the cyclooxygenases of teleost fishes. Mol. Phylogen. Evol. 56:451-455.

Henschel R, Lieber M, Wu L-S, Nista PM, Haas BJ, LeDuc RD. 2012. Trinity RNA-Seq assembler performance optimization. Proceedings of the 1st Conference of the Extreme Science and Engineering Discovery Environment: Bridging from the eXtreme to the campus and beyond. Chicago, Illinois, USA, ACM, p. 1-8.

Hillis DM. 1995. Approaches for Assessing Phylogenetic Accuracy. Syst. Biol. 44:3-16.

Hillis DM, Pollock DD, McGuire JA, Zwickl DJ. 2003. Is Sparse Taxon Sampling a Problem for Phylogenetic Inference? Syst. Biol. 52:124-126.

Hittinger CT, Johnston M, Tossberg JT, Rokas A. 2010. Leveraging skewed transcript abundance by RNA-Seq to increase the genomic depth of the tree of life. Proc. Natl. Acad. Sci. USA 107:1476-1481.

Honaas LA, Wafula EK, Wickett NJ, Der JP, Zhang Y, Edger PP, Altman NS, Pires JC, Leebens-Mack JH, dePamphilis CW. 2016. Selecting Superior *De Novo* Transcriptome Assemblies: Lessons Learned by Leveraging the Best Plant Genome. PLoS ONE 11:e0146062.

Howison M, Sinnott-Armstrong NA, Dunn CW. 2012. BioLite, a lightweight bioinformatics framework with automated tracking of diagnostics and provenance. Proceedings of the 4th USENIX Workshop on the Theory and Practice of Provenance (TaPP '12). Boston, MA, USA.

Huelsenbeck JP. 1995. Performance of Phylogenetic Methods in Simulation. Syst. Biol. 44:17-48.

Huerta-Cepas J, Serra F, Bork P. 2016. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. Mol. Biol. Evol. 33:1635-1638.

Janouškovec J, Gavelis GS, Burki F, Dinh D, Bachvaroff TR, Gornik SG, Bright KJ, Imanian B, Strom SL, Delwiche CF, Waller RF, Fensome RA, Leander BS, Rohwer FL, Saldarriaga JF. 2017. Major transitions in dinoflagellate evolution unveiled by phylotranscriptomics. Proc. Natl. Acad. Sci. USA 114:E171-E180.

Jarvis ED, Mirarab S, Aberer AJ, Li B, Houde P, Li C, Ho SYW, Faircloth BC, Nabholz B, Howard JT, Suh A, Weber CC, da Fonseca RR, Li J, Zhang F, Li H, Zhou L, Narula N, Liu L, Ganapathy G, Boussau B, Bayzid MS, Zavidovych V, Subramanian S, Gabaldón T, Capella-Gutiérrez S, Huerta-Cepas J, Rekepalli B, Munch K, Schierup M, Lindow B, Warren WC, Ray D, Green RE, Bruford MW, Zhan X, Dixon A, Li S, Li N, Huang Y, Derryberry EP, Bertelsen MF, Sheldon FH, Brumfield RT, Mello CV, Lovell PV, Wirthlin M, Schneider MPC, Prosdocimi F, Samaniego JA, Velazquez AMV, Alfaro-Núñez A, Campos PF, Petersen B, Sicheritz-Ponten T, Pas A, Bailey T, Scofield P, Bunce M, Lambert DM, Zhou Q, Perelman P, Driskell AC, Shapiro B, Xiong Z, Zeng Y, Liu S, Li Z, Liu B, Wu K, Xiao J, Yinqi X, Zheng Q, Zhang Y, Yang H, Wang J, Smeds L, Rheindt FE, Braun M, Fjeldsa J, Orlando L, Barker FK, Jønsson KA, Johnson W, Koepfli K-P, O'Brien S, Haussler D, Ryder OA, Rahbek C, Willerslev E, Graves GR, Glenn TC, McCormack J, Burt D, Ellegren H, Alström P, Edwards SV, Stamatakis A, Mindell DP, Cracraft J, Braun EL, Warnow T, Jun W, Gilbert MTP, Zhang G. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. Science 346:1320-1331.

Jun J, Mandoiu II, Nelson CE. 2009. Identification of mammalian orthologs using local synteny. BMC Genomics 10:630.

Katoh K, Misawa K, Kuma Ki, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30:3059-3066.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30:772-780.

Kobert K, Salichos L, Rokas A, Stamatakis A. 2016. Computing the Internode Certainty and related measures from partial gene trees. Mol. Biol. Evol. 33:1606-1617.

Lamichhaney S, Berglund J, Almen MS, Maqbool K, Grabherr M, Martinez-Barrio A, Promerova M, Rubin C-J, Wang C, Zamani N, Grant BR, Grant PR, Webster MT, Andersson L. 2015. Evolution of Darwin's finches and their beaks revealed by genome sequencing. Nature 518:371-375.

Lechner M, Hernandez-Rosales M, Doerr D, Wieseke N, Thévenin A, Stoye J, Hartmann RK, Prohaska SJ, Stadler PF. 2014. Orthology Detection Combining Clustering and Synteny for Very Large Datasets. PLoS ONE 9:e105015.

Lee EK, Cibrian-Jaramillo A, Kolokotronis S-O, Katari MS, Stamatakis A, Ott M, Chiu JC, Little DP, Stevenson DW, McCombie WR, Martienssen RA, Coruzzi G, DeSalle R. 2011. A Functional Phylogenomic View of the Seed Plants. PLoS Genet. 7:e1002411.

Lei M, Dong D. 2016. Phylogenomic analyses of bat subordinal relationships based on transcriptome data. Scientific Reports 6:27726.

Lemmon AR, Emme SA, Lemmon EM. 2012. Anchored Hybrid Enrichment for Massively High-Throughput Phylogenomics. Syst. Biol. 61:727-744.

Lemmon EM, Lemmon AR. 2013. High-Throughput Genomic Data in Systematics and Phylogenetics. Annu. Rev. Ecol., Evol. Syst. 44:99-121.

Li L, Stoeckert CJ, Roos DS. 2003. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. Genome Res. 13:2178-2189.

Librado P, Der Sarkissian C, Ermini L, Schubert M, Jónsson H, Albrechtsen A, Fumagalli M, Yang MA, Gamba C, Seguin-Orlando A, Mortensen CD, Petersen B, Hoover CA, Lorente-Galdos B, Nedoluzhko A, Boulygina E, Tsygankova S, Neuditschko M, Jagannathan V, Thèves C, Alfarhan AH, Alquraishi SA, Al-Rasheid KAS, Sicheritz-Ponten T, Popov R, Grigoriev S, Alekseev AN, Rubin EM, McCue M, Rieder S, Leeb T, Tikhonov A, Crubézy E, Slatkin M, Marques-Bonet T, Nielsen R, Willerslev E, Kantanen J, Prokhortchouk E, Orlando L. 2015. Tracking the origins of Yakutian horses and the genetic basis for their fast adaptation to subarctic environments. Proc. Natl. Acad. Sci. USA 112:E6889-E6897.

Lin Q, Fan S, Zhang Y, Xu M, Zhang H, Yang Y, Lee AP, Woltering JM, Ravi V, Gunter HM, Luo W, Gao Z, Lim ZW, Qin G, Schneider RF, Wang X, Xiong P, Li G, Wang K, Min J, Zhang C, Qiu Y, Bai J, He W, Bian C, Zhang X, Shan D, Qu H, Sun Y, Gao Q, Huang L, Shi Q, Meyer A, Venkatesh B. 2016. The seahorse genome and the evolution of its specialized morphology. Nature 540:395-399.

Lyons E, Freeling M. 2008. How to usefully compare homologous plant genes and chromosomes as DNA sequences. Plant J. 53:661-673.

Lyons E, Pedersen B, Kane J, Freeling M. 2008. The Value of Nonmodel Genomes and an Example Using SynMap Within CoGe to Dissect the Hexaploidy that Predates the Rosids. Tropical Plant Biol. 1:181-190.

Malinsky M, Challis RJ, Tyers AM, Schiffels S, Terai Y, Ngatunga BP, Miska EA, Durbin R, Genner MJ, Turner GF. 2015. Genomic islands of speciation separate cichlid ecomorphs in an East African crater lake. Science 350:1493-1498.

McCormack JE, Tsai WLE, Faircloth BC. 2016. Sequence capture of ultraconserved elements from bird museum specimens. Molecular Ecology Resources 16:1189-1203.

McKain MR, Tang H, McNeal JR, Ayyampalayam S, Davis JI, dePamphilis CW, Givnish TJ, Pires JC, Stevenson DW, Leebens-Mack JH. 2016. A Phylogenomic Assessment of Ancient Polyploidy and Genome Evolution across the Poales. GBE 8:1150-1164.

McKain MR, Wickett N, Zhang Y, Ayyampalayam S, McCombie WR, Chase MW, Pires JC, dePamphilis CW, Leebens-Mack J. 2012. Phylogenomic analysis of transcriptome data elucidates co-occurrence of a paleopolyploid event and the origin of bimodal karyotypes in Agavoideae (Asparagaceae). Am. J. Bot. 99:397-406.

Mirarab S, Bayzid MS, Boussau B, Warnow T. 2014a. Statistical binning enables an accurate coalescent-based estimation of the avian tree. Science 346:1250463.

Mirarab S, Reaz R, Bayzid MS, Zimmermann T, Swenson MS, Warnow T. 2014b. ASTRAL: genome-scale coalescent-based species tree estimation. Bioinformatics 30:i541-i548.

Misof B, Liu S, Meusemann K, Peters RS, Donath A, Mayer C, Frandsen PB, Ware J, Flouri T, Beutel RG, Niehuis O, Petersen M, Izquierdo-Carrasco F, Wappler T, Rust J, Aberer AJ, Aspöck U, Aspöck H, Bartel D, Blanke A, Berger S, Böhm A, Buckley TR, Calcott B, Chen J, Friedrich F, Fukui M, Fujita M, Greve C, Grobe P, Gu S, Huang Y, Jermiin LS, Kawahara AY, Krogmann L, Kubiak M, Lanfear R, Letsch H, Li Y, Li Z, Li J, Lu H, Machida R, Mashimo Y, Kapli P, McKenna DD, Meng G, Nakagaki Y, Navarrete-Heredia JL, Ott M, Ou Y, Pass G, Podsiadlowski L, Pohl H, von Reumont BM, Schütte K, Sekiya K, Shimizu S, Slipinski A, Stamatakis A, Song W, Su X, Szucsich NU, Tan M, Tan X, Tang M,

Tang J, Timelthaler G, Tomizuka S, Trautwein M, Tong X, Uchifune T, Walzl MG, Wiegmann BM, Wilbrandt J, Wipfler B, Wong TKF, Wu Q, Wu G, Xie Y, Yang S, Yang Q, Yeates DK, Yoshizawa K, Zhang Q, Zhang R, Zhang W, Zhang Y, Zhao J, Zhou C, Zhou L, Ziesmann T, Zou S, Li Y, Xu X, Zhang Y, Yang H, Wang J, Wang J, Kjer KM, Zhou X. 2014. Phylogenomics resolves the timing and pattern of insect evolution. Science 346:763-767.

Moyle RG, Oliveros CH, Andersen MJ, Hosner PA, Benz BW, Manthey JD, Travers SL, Brown RM, Faircloth BC. 2016. Tectonic collision and uplift of Wallacea triggered the global songbird radiation. Nature Communications 7:12709.

Nabhan AR, Sarkar IN. 2011. The impact of taxon sampling on phylogenetic inference: a review of two decades of controversy. Briefings in Bioinformatics 13:122-134.

Neafsey DE, Waterhouse RM, Abai MR, Aganezov SS, Alekseyev MA, Allen JE, Amon J, Arcà B, Arensburger P, Artemov G, Assour LA, Basseri H, Berlin A, Birren BW, Blandin SA, Brockman AI, Burkot TR, Burt A, Chan CS, Chauve C, Chiu JC, Christensen M, Costantini C, Davidson VLM, Deligianni E, Dottorini T, Dritsou V, Gabriel SB, Guelbeogo WM, Hall AB, Han MV, Hlaing T, Hughes DST, Jenkins AM, Jiang X, Jungreis I, Kakani EG, Kamali M, Kemppainen P, Kennedy RC, Kirmitzoglou IK, Koekemoer LL, Laban N, Langridge N, Lawniczak MKN, Lirakis M, Lobo NF, Lowy E, MacCallum RM, Mao C, Maslen G, Mbogo C, McCarthy J, Michel K, Mitchell SN, Moore W, Murphy KA, Naumenko AN, Nolan T, Novoa EM, O'Loughlin S, Oringanje C, Oshaghi MA, Pakpour N, Papathanos PA, Peery AN, Povelones M, Prakash A, Price DP, Rajaraman A, Reimer LJ, Rinker DC, Rokas A, Russell TL, Sagnon NF, Sharakhova MV, Shea T, Simão FA, Simard F, Slotman MA, Somboon P, Stegniy V, Struchiner CJ, Thomas GWC, Tojo M, Topalis P, Tubio JMC, Unger MF, Vontas J, Walton C, Wilding CS, Willis JH, Wu Y-C, Yan G, Zdobnov EM, Zhou X, Catteruccia F, Christophides GK, Collins FH, Cornman RS, Crisanti A, Donnelly MJ, Emrich SJ, Fontaine MC, Gelbart W, Hahn MW, Hansen IA, Howell PI, Kafatos FC, Kellis M, Lawson D, Louis C, Luckhart S, Muskavitch MAT, Ribeiro JM, Riehle MA, Sharakhov IV, Tu Z, Zwiebel LJ, Besansky NJ. 2015. Highly evolvable malaria vectors: The genomes of 16 *Anopheles* mosquitoes. Science 347.

Orlando L, Ginolhac A, Zhang G, Froese D, Albrechtsen A, Stiller M, Schubert M, Cappellini E, Petersen B, Moltke I, Johnson PLF, Fumagalli M, Vilstrup JT, Raghavan M, Korneliussen T, Malaspinas A-S, Vogt J, Szklarczyk D, Kelstrup CD, Vinther J, Dolocan A, Stenderup J, Velazquez AMV, Cahill J, Rasmussen M, Wang X, Min J, Zazula GD, Seguin-Orlando A, Mortensen C, Magnussen K, Thompson JF, Weinstock J, Gregersen K, Roed KH, Eisenmann V, Rubin CJ, Miller DC, Antczak DF, Bertelsen MF, Brunak S, Al-Rasheid KAS, Ryder O, Andersson L, Mundy J, Krogh A, Gilbert MTP, Kjaer K, Sicheritz-Ponten T, Jensen LJ, Olsen JV, Hofreiter M, Nielsen R, Shapiro B, Wang J, Willerslev E. 2013. Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse. Nature 499:74-78.

Ouyang S, Zhu W, Hamilton J, Lin H, Campbell M, Childs K, Thibaud-Nissen F, Malek RL, Lee Y, Zheng L, Orvis J, Haas B, Wortman J, Buell CR. 2007. The TIGR Rice Genome Annotation Resource: improvements and new features. Nucleic Acids Res. 35:D883-D887.

Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A, Schmutz J, Spannagl M, Tang H, Wang X, Wicker T, Bharti AK, Chapman J, Feltus FA, Gowik U, Grigoriev IV, Lyons E, Maher CA, Martis M, Narechania A, Otillar RP, Penning BW, Salamov AA, Wang Y, Zhang L, Carpita NC, Freeling M, Gingle AR, Hash CT, Keller B, Klein P, Kresovich S, McCann MC, Ming R, Peterson DG, Mehboob ur R, Ware D, Westhoff P, Mayer KFX, Messing J, Rokhsar DS. 2009. The Sorghum bicolor genome and the diversification of grasses. Nature 457:551-556.

Pease JB, Haak DC, Hahn MW, Moyle LC. 2016. Phylogenomics Reveals Three Sources of Adaptive Variation during a Rapid Radiation. PLoS Biol. 14:e1002379.

Prasanna AN, Mehra S. 2013. Comparative Phylogenomics of Pathogenic and Non-Pathogenic Mycobacterium. PLOS ONE 8:e71248.

Salichos L, Rokas A. 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. Nature 497:327-331.

Salichos L, Stamatakis A, Rokas A. 2014. Novel Information Theory-Based Measures for Quantifying Incongruence among Phylogenetic Trees. Mol. Biol. Evol. 31:1261-1271.

Schmickl R, Liston A, Zeisek V, Oberlander K, Weitemier K, Straub SCK, Cronn RC, Dreyer LL, Suda J. 2016. Phylogenetic marker development for target enrichment from transcriptome and genome skim data: the pipeline and its application in southern African Oxalis (Oxalidaceae). Molecular Ecology Resources 16:1124-1135.

Schmieder R, Edwards R. 2011. Quality control and preprocessing of metagenomic datasets. Bioinformatics 27:863-864.

Schnable JC, Freeling M, Lyons E. 2012a. Genome-wide analysis of syntenic gene deletion in the grasses. Genome Biol Evol 4:265-277.

Schnable JC, Springer NM, Freeling M. 2011. Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. Proc. Natl. Acad. Sci. USA 108:4069-4074.

Schnable JC, Wang X, Pires JC, Freeling M. 2012b. Escape from preferential retention following repeated whole genome duplications in plants. Front. Plant Sci. 3:94.

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA, Minx P, Reily AD, Courtney L, Kruchowski SS, Tomlinson C,

Strong C, Delehaunty K, Fronick C, Courtney B, Rock SM, Belter E, Du F, Kim K, Abbott RM, Cotton M, Levy A, Marchetto P, Ochoa K, Jackson SM, Gillam B, Chen W, Yan L, Higginbotham J, Cardenas M, Waligorski J, Applebaum E, Phelps L, Falcone J, Kanchi K, Thane T, Scimone A, Thane N, Henke J, Wang T, Ruppert J, Shah N, Rotter K, Hodges J, Ingenthron E, Cordes M, Kohlberg S, Sgro J, Delgado B, Mead K, Chinwalla A, Leonard S, Crouse K, Collura K, Kudrna D, Currie J, He R, Angelova A, Rajasekar S, Mueller T, Lomeli R, Scara G, Ko A, Delaney K, Wissotski M, Lopez G, Campos D, Braidotti M, Ashley E, Golser W, Kim H, Lee S, Lin J, Dujmic Z, Kim W, Talag J, Zuccolo A, Fan C, Sebastian A, Kramer M, Spiegel L, Nascimento L, Zutavern T, Miller B, Ambroise C, Muller S, Spooner W, Narechania A, Ren L, Wei S, Kumari S, Faga B, Levy MJ, McMahan L, Van Buren P, Vaughn MW, Ying K, Yeh C-T, Emrich SJ, Jia Y, Kalyanaraman A, Hsia A-P, Barbazuk WB, Baucom RS, Brutnell TP, Carpita NC, Chaparro C, Chia J-M, Deragon J-M, Estill JC, Fu Y, Jeddeloh JA, Han Y, Lee H, Li P, Lisch DR, Liu S, Liu Z, Nagel DH, McCann MC, SanMiguel P, Myers AM, Nettleton D, Nguyen J, Penning BW, Ponnala L, Schneider KL, Schwartz DC, Sharma A, Soderlund C, Springer NM, Sun Q, Wang H, Waterman M, Westerman R, Wolfgruber TK, Yang L, Yu Y, Zhang L, Zhou S, Zhu Q, Bennetzen JL, Dawe RK, Jiang J, Jiang N, Presting GG, Wessler SR, Aluru S, Martienssen RA, Clifton SW, McCombie WR, Wing RA, Wilson RK. 2009. The B73 Maize Genome: Complexity, Diversity, and Dynamics. Science 326:1112-1115.

Smith SA, Moore MJ, Brown JW, Yang Y. 2015. Analysis of phylogenomic datasets reveals conflict, concordance, and gene duplications with examples from animals and plants. BMC Evol. Biol. 15:1-15.

Smith SA, Pease JB. 2016. Heterogeneous molecular processes among the causes of how sequence similarity scores can fail to recapitulate phylogeny. Brief Bioinform:bbw034.

Spriggs EL, Christin P-A, Edwards EJ. 2014. C4 Photosynthesis Promoted Species Diversification during the Miocene Grassland Expansion. PLoS ONE 9:e97722.

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:2688-2690.

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

Stevens PF. 2017. Angiosperm Phylogeny Website.

Studer AJ, Schnable JC, Weissmann S, Kolbe AR, McKain MR, Shao Y, Cousins AB, Kellogg EA, Brutnell TP. 2016. The draft genome of the C3 panicoid grass species Dichanthelium oligosanthes. Genome Biol. 17:223.

Sveinsson S, McDill J, Wong GKS, Li J, Li X, Deyholos MK, Cronk QCB. 2014. Phylogenetic pinpointing of a paleopolyploidy event within the flax genus (*Linum*) using transcriptomics. Ann. Bot. 113:753-761.

Tang H, Bowers JE, Wang X, Ming R, Alam M, Paterson AH. 2008. Synteny and Collinearity in Plant Genomes. Science 320:486-488.

Tang H, Lyons E, Pedersen B, Schnable JC, Paterson AH, Freeling M. 2011. Screening synteny blocks in pairwise genome comparisons through integer programming. BMC Bioinformatics 12:1-11.

The International Brachypodium Initiative. 2010. Genome sequencing and analysis of the model grass Brachypodium distachyon. Nature 463:763-768.

Todd EV, Black MA, Gemmell NJ. 2016. The power and promise of RNA-seq in ecology and evolution. Mol. Ecol. (Online ahead of print).

Tsagkogeorga G, Parker J, Stupka E, Cotton JA, Rossiter SJ. 2013. Phylogenomic Analyses Elucidate the Evolutionary Relationships of Bats. Curr. Biol. 23:2262-2267.

van Dongen S. 2000. Graph Clustering by Flow Simulation. University of Utrecht.

VanBuren R, Bryant D, Edger PP, Tang H, Burgess D, Challabathula D, Spittle K, Hall R, Gu J, Lyons E, Freeling M, Bartels D, Ten Hallers B, Hastie A, Michael TP, Mockler TC. 2015. Single-molecule sequencing of the desiccation-tolerant grass Oropetium thomaeum. Nature 527:508-511.

Vicentini A, Barber JC, Aliscioni SS, Giussani LM, Kellogg EA. 2008. The age of the grasses and clusters of origins of C4 photosynthesis. Global Change Biol. 14:2963-2977.

Wang L, Si Y, Dedow LK, Shao Y, Liu P, Brutnell TP. 2011. A Low-Cost Library Construction Protocol and Data Analysis Pipeline for Illumina-Based Strand-Specific Multiplex RNA-Seq. PLoS ONE 6:e26426.

Wang Z, Wu M. 2015. An integrated phylogenomic approach toward pinpointing the origin of mitochondria. Scientific Reports 5:7949.

Washburn JD, Bird KA, Conant GC, Pires JC. 2016. Convergent Evolution and the Origin of Complex Phenotypes in the Age of Systems Biology. Int. J. Plant Sci. 177:305-318.

Washburn JD, Schnable JC, Davidse G, Pires JC. 2015. Phylogeny and photosynthesis of the grass tribe Paniceae. Am. J. Bot. 102:1493-1505.

Weitemier K, Straub SCK, Cronn RC, Fishbein M, Schmickl R, McDonnell A, Liston A. 2014. Hyb-Seq: Combining Target Enrichment and Genome Skimming for Plant Phylogenomics. Appl. Plant Sci. 2:1400042.

Wen J, Xiong Z, Nie Z-L, Mao L, Zhu Y, Kan X-Z, Ickert-Bond SM, Gerrath J, Zimmer EA, Fang X-D. 2013. Transcriptome Sequences Resolve Deep Relationships of the Grape Family. PLOS ONE 8:e74394.

Wickett NJ, Honaas LA, Wafula EK, Das M, Huang K, Wu B, Landherr L, Timko MP, Yoder J, Westwood JH, dePamphilis CW. 2011. Transcriptomes of the Parasitic Plant Family Orobanchaceae Reveal Surprising Conservation of Chlorophyll Synthesis. Curr. Biol. 21:2098-2104.

Wickett NJ, Mirarab S, Nguyen N, Warnow T, Carpenter E, Matasci N, Ayyampalayam S, Barker MS, Burleigh JG, Gitzendanner MA, Ruhfel BR, Wafula E, Der JP, Graham SW, Mathews S, Melkonian M, Soltis DE, Soltis PS, Miles NW, Rothfels CJ, Pokorny L, Shaw AJ, DeGironimo L, Stevenson DW, Surek B, Villarreal JC, Roure B, Philippe H, dePamphilis CW, Chen T, Deyholos MK, Baucom RS, Kutchan TM, Augustin MM, Wang J, Zhang Y, Tian Z, Yan Z, Wu X, Sun X, Wong GK-S, Leebens-Mack J. 2014. Phylotranscriptomic analysis of the origin and early diversification of land plants. Proc. Natl. Acad. Sci. USA 111:E4859-E4868.

Xi Z, Liu L, Rest JS, Davis CC. 2014. Coalescent versus Concatenation Methods and the Placement of Amborella as Sister to Water Lilies. Syst. Biol. 63:919-932.

Yang Y, Moore MJ, Brockington SF, Soltis DE, Wong GK-S, Carpenter EJ, Zhang Y, Chen L, Yan Z, Xie Y, Sage RF, Covshoff S, Hibberd JM, Nelson MN, Smith SA. 2015a. Dissecting Molecular Evolution in the Highly Diverse Plant Clade Caryophyllales Using Transcriptome Sequencing. Mol. Biol. Evol. 32:2001-2014.

Yang Y, Moore MJ, Brockington SF, Timoneda A, Feng T, Marx HE, Walker JF, Smith SA. 2017. An Efficient Field and Laboratory Workflow for Plant Phylotranscriptomic Projects. Appl. Plant Sci. 5:1600128.

Yang Y, Smith S. 2013. Optimizing de novo assembly of short-read RNA-seq data for phylogenomics. BMC Genomics 14:328.

Yang Y, Smith SA. 2014. Orthology Inference in Nonmodel Organisms Using Transcriptomes and Low-Coverage Genomes: Improving Accuracy and Matrix Occupancy for Phylogenomics. Mol. Biol. Evol. 31:3081-3092.

Yang Z, Wafula EK, Honaas LA, Zhang H, Das M, Fernandez-Aparicio M, Huang K, Bandaranayake PCG, Wu B, Der JP, Clarke CR, Ralph PE, Landherr L, Altman NS, Timko MP, Yoder JI, Westwood JH, dePamphilis CW. 2015b. Comparative Transcriptome Analyses Reveal Core Parasitism Genes and Suggest Gene Duplication and Repurposing as Sources of Structural Novelty. Mol. Biol. Evol. 32:767-790.

Zhang G, Li C, Li Q, Li B, Larkin DM, Lee C, Storz JF, Antunes A, Greenwold MJ, Meredith RW, Ödeen A, Cui J, Zhou Q, Xu L, Pan H, Wang Z, Jin L, Zhang P, Hu H, Yang W, Hu J, Xiao J, Yang Z, Liu Y, Xie Q, Yu H, Lian J, Wen P, Zhang F, Li H, Zeng Y, Xiong Z, Liu S, Zhou L, Huang Z, An N, Wang J, Zheng Q, Xiong Y, Wang G, Wang B, Wang J, Fan Y, da Fonseca RR, Alfaro-Núñez A, Schubert M, Orlando L, Mourier T, Howard JT, Ganapathy G, Pfenning A, Whitney O, Rivas MV, Hara E, Smith J, Farré M, Narayan J, Slavov G, Romanov MN, Borges R, Machado JP, Khan I, Springer MS, Gatesy J, Hoffmann FG, Opazo JC, Håstad O, Sawyer RH, Kim H, Kim K-W, Kim HJ, Cho S, Li N, Huang Y, Bruford MW, Zhan X, Dixon A, Bertelsen MF, Derryberry E, Warren W, Wilson RK, Li S, Ray DA, Green RE, O'Brien SJ, Griffin D, Johnson WE, Haussler D, Ryder OA, Willerslev E, Graves GR, Alström P, Fjeldså J, Mindell DP, Edwards SV, Braun EL, Rahbek C, Burt DW, Houde P, Zhang Y, Yang H, Wang J, Consortium AG, Jarvis ED, Gilbert MTP, Wang J. 2014. Comparative genomics reveals insights into avian genome evolution and adaptation. Science 346:1311-1320.

Zimmer EA, Wen J. 2015. Using nuclear gene data for plant phylogenetics: Progress and prospects II. Next-gen approaches. J. Syst. Evol. 53:371-379.

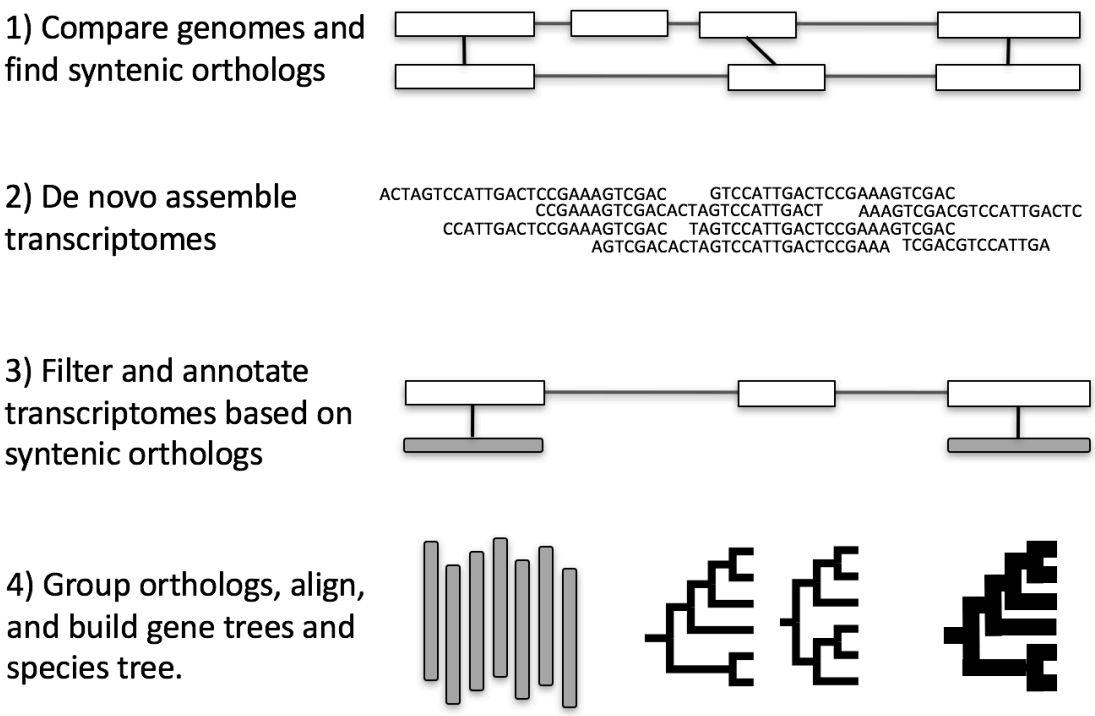


Figure 3.1 - Genome-guided phylo-transcriptomics workflow.
Illustration of the workflow followed to produce the genome-guided phylogenies in this study.

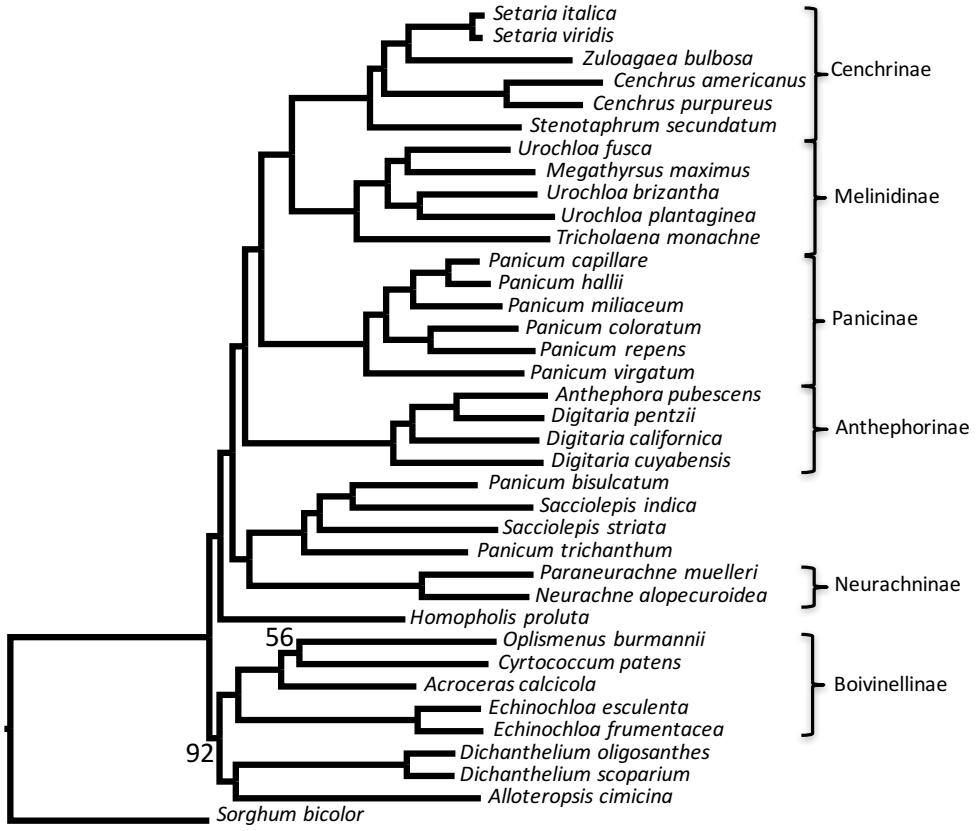


Figure 3.2 - Genome-guided concatenation-based phylogeny of the tribe Paniceae. Phylogenetic tree of the tribe Paniceae (Poaceae) built using RAxML based on a concatenated matrix with 90% gene occupancy. Branches are labeled with maximum likelihood bootstrap values; unlabeled branches have values of 100.

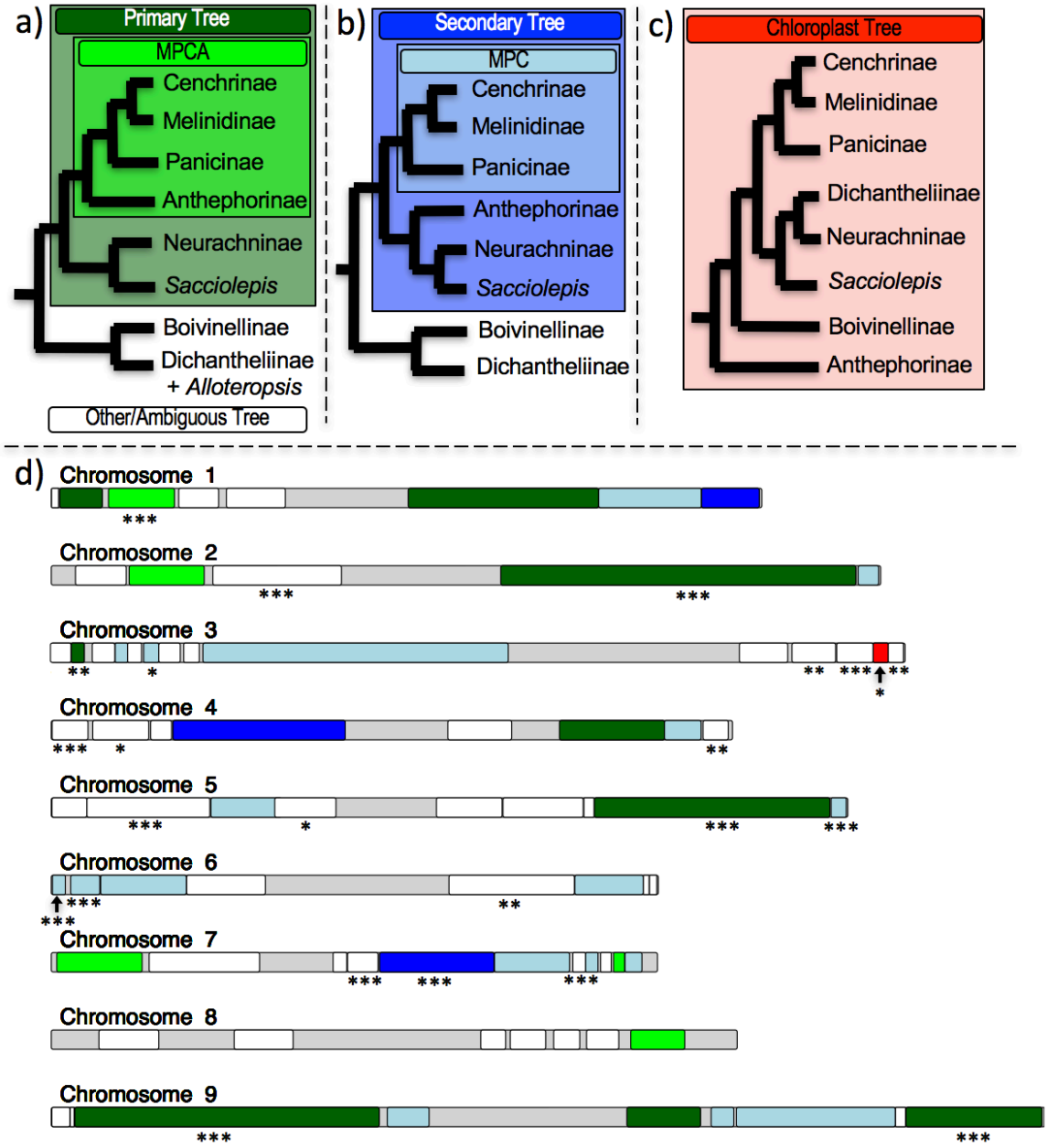


Figure 3.3 - Phylogenies Mapped to Chromosome Blocks. a) Primary nuclear topology found using all methods, b) Secondary nuclear topology, c) Chloroplast topology re-drawn from Washburn, et al. (2015). d) An ideogram of the *Setaria italica* chromosomes with conserved syntenic blocks between *S. italica* and *Sorghum bicolor* demarcated. Syntenic blocks are colored based on the phylogenetic patterns from a-c that each block supports. Gray indicates areas of the chromosomes not covered by our blocks. Asterisks below the blocks indicate significance level for pairwise Robinson-Foulds distance tests: *** <0.001, ** <0.01, *<0.05.

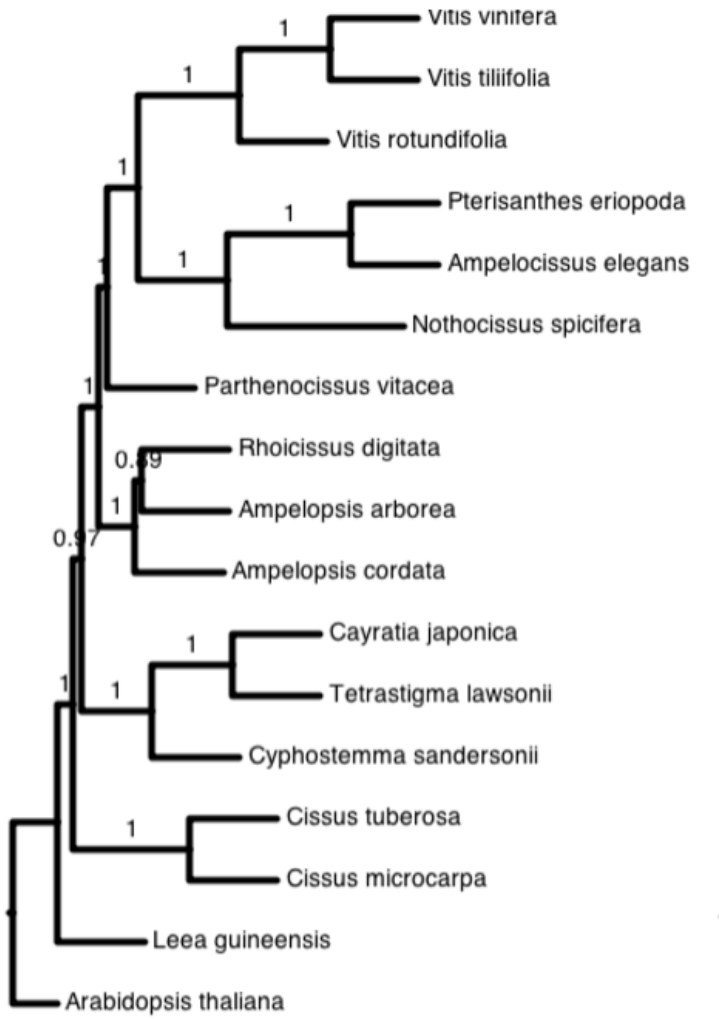


Figure 3.4 - Tree built using the Grape data based on the genome-guided method. Trees from all three methods shared this same topology.

Table 3.1 - Total orthologs found in each method separated by matrix occupancy.

Method			8 spp	90%	Full
Genome-guided	Genes	Total	9,757	2,211	434
		Min	5,389	1,963	434
	Amino Acids	Total	4,182,364	835,229	144,503
		Min	1,775,925	669,215	128,896
Agalma	Genes	Total	11,563	2,308	555
		Min	5,453	2,054	555
	Amino Acids	Total	4,420,707	797,333	182,368
		Min	1,568,329	613,538	168,157
Yang & Smith 1 to 1	Genes	Total	7,323	1,925	898
		Min	3,685	1,781	898
	Amino Acids	Total	2,408,802	789,203	361,901
		Min	1,129,993	628,190	310,283
Yang & Smith MO	Genes	Total	11,568	1,966	1,076
		Min	6,417	1,879	1,076
	Amino Acids	Total	4,362,686	857,857	456,597
		Min	2,009,430	687,942	380,988

Table 3.2 - Approximate run times in hours (hrs) for each orthology inference method based on a 16 CPU system.

	Synteny Step	BLAST Step	Alignment and Tree Building for Pruning	Total
Genome-Guided	< 1	6.7	N/A	7.7
Agalma	N/A	46.4	88.6	135.0
Yang & Smith	N/A	366.9	412.2	779.1

Table 3.3 - Grass (Poaceae) wide gene by gene comparisons of orthology detection methods to a benchmark set of orthologs derived entirely from syntenic relationships between sequenced genomes.

Method			4 species	5 species	6 species
Genome-Guided	All Trees Included	Trees Agreeing with Benchmark	4,119	2,169	413
		Total Trees	6,669	3,700	896
		Percent Trees in Agreement	61.8%	58.6%	46.1%
Yang & Smith 1 to 1	All Trees Included	Trees Agreeing with Benchmark	1,936	1,741	1,370
		Total Trees	7,933	6,989	5,171
		Percent Trees in Agreement	24.4%	24.9%	26.5%
	Excluding trees not in benchmark set	Trees Agreeing with Benchmark	1,936	1,741	1,370
		Total Trees	6,088	5,417	4,320
		Percent Trees in Agreement	31.8%	32.1%	31.7%
Yang & Smith MO	All Trees Included	Trees Agreeing with Benchmark	2,000	1,795	1,404
		Total Trees	8,619	7,560	5,464
		Percent Trees in Agreement	23.2%	23.7%	25.7%
	Excluding trees not in benchmark set	Trees Agreeing with Benchmark	2,000	1,795	1,404
		Total Trees	6,503	5,757	4,516
		Percent Trees in Agreement	30.8%	31.2%	31.1%

Supplemental Table S3.1 - Materials used in study.

A list of all plant species used in the study along with their source, identification number, herbarium specimen number, and NCBI record numbers where applicable.

Subtribe	Genus	Species	Authority	Source	ID number	Herbarium Accession No.	NCBI Number	Library Method
Cenchrinae	*Cenchrus*	*americanus*	(L.) Morrone	KD	ICMP-451	MO-6635001	XXXX	TS
Cenchrinae	*Cenchrus*	*americanus*	(L.) Morrone	USDA	PI 279664	N/A	XXXX	W
Cenchrinae	*Cenchrus*	*purpureus*	(Schumach.) Morrone	USDA	PI 667860	N/A	XXXX	TS
Cenchrinae	*Setaria*	*viridis*	(L.) P. Beauv.	N/A	N/A	N/A	ERR385861-6	N/A
Cenchrinae	*Stenotaphrum*	*secundatum*	(Walter) Kuntze	USDA	PI 410357	MO-6635005	XXXX	TS
Cenchrinae	*Zuloagaea*	*bulbosa*	(Kunth) Bess	USDA	PI 442528	MO-6635006	XXXX	TS
Melinidinae	*Megathyrsus*	*maximus*	(Jacq.) B.K. Simon & S.W.L. Jacobs	USDA	PI 404634	MO-6635008	XXXX	TS
Melinidinae	*Tricholaena*	*monachne*	(Trin.) Stapf & C.E. Hubb.	USDA	PI 166381	MO-6635009	XXXX	TS
Melinidinae	*Urochloa*	*brizantha*	(Hochst. ex A. Rich.) R. Webster	USDA	PI 226049	MO-6635010	XXXX	W
Melinidinae	*Urochloa*	*fusca*	(Sw.) B.F. Hansen & Wunderlin	USDA	LBJWC-52	MO-6635011	XXXX	W
Melinidinae	*Urochloa*	*plantaginea*	(Link) R.D. Webster	USDA	PI 379628	MO-6635012	XXXX	TS
Panicinae	*Panicum*	*capillare*	L.	USDA	PI 220025	MO-6635013	XXXX	TS
Panicinae	*Panicum*	*coloratum*	L.	USDA	PI 185546	MO-6635014	XXXX	TS
Panicinae	*Panicum*	*hallii*	Vasey	DL	HAL 2	MO-6635015	XXXX	TS
Panicinae	*Panicum*	*miliaceum*	L.	USDA	PI 578073	MO-6635016	XXXX	W
Panicinae	*Panicum*	*repens*	L.	USDA	PI 208687	MO-6635017	XXXX	TS
Panicinae	*Panicum*	*repens*	L.	USDA	PI 238344	MO-6635018	XXXX	TS
Panicinae	*Panicum*	*virgatum*	L.	LEB	AP13	MO-6635019	XXXX	TS
Dichantheliinae	*Dichanthelium*	*oligosanthes*	(Schult.) Gould	AS	D1	MO-6635020	XXXX	W
Dichantheliinae	*Dichanthelium*	*scoparium*	(Lam.) Gould	USDA	PI 652864	MO-6635021	XXXX	TS
Incertae sedis	*Panicum*	*bisulcatum*	Thunb.	USDA	PI 286485	MO-6647157	XXXX	TS
Incertae sedis	*Panicum*	*trichanthum*	Nees	USDA	PI 206329	MO-6647158	XXXX	TS
Incertae sedis	*Sacciolepis*	*indica*	(L.) Chase	USDA	PI 338609	MO-6635022	XXXX	TS
Incertae sedis	*Sacciolepis*	*striata*	(L.) Nash	USDA	NSL 454620	MO-6635023	XXXX	TS
Incertae sedis	*Walwhalleya*	*proluta*	(F. Muell.) K. E. Wills & J. J. Bruhl	NS	NS 42146	MO-6647159	XXXX	TS
Neurachninae	*Neurachne*	*alopecuroidea*	R. Br.	ML	N/A	N/A	XXXX	TS
Neurachninae	*Paraneurachne*	*muelleri*	(Hack.) S.T.Blake	ML	N/A	N/A	XXXX	TS
Boivinellinae	*Acroceras*	*calcicola*	A. Camus	MSB	MSB 199378	MO-6647161	XXXX	TS
Boivinellinae	*Alloteropsis*	*cimicina*	(L.) Stapf	JRB	JRB	MO-6635025	XXXX	TS
Boivinellinae	*Cyrtococcum*	*patens*	(L.) A. Camus	MSB	MSB 516	MO-6647160	XXXX	TS
Boivinellinae	*Echinochloa*	*esculenta*	(A. Braun) H. Scholz	USDA	PI 647850	MO-6635026	XXXX	W
Boivinellinae	*Echinochloa*	*frumentacea*	Link	USDA	Ames 11429	MO-6635027	XXXX	W
Boivinellinae	*Oplismenus*	*burmannii*	(Retz.) P. Beauv.	MW	MW	MO-6635028	XXXX	TS
Anthephorinae	*Anthephora*	*pubescens*	Nees	EK	TK1	MO-6635029	XXXX	W
Anthephorinae	*Digitaria*	*californica*	(Benth.) Henrard	USDA	PI 364670	MO-6635030	XXXX	TS
Anthephorinae	*Digitaria*	*cuyabensis*	(Trin.) Parodi	USDA	PI 349688	MO-6635031	XXXX	TS
Anthephorinae	*Digitaria*	*pentzii*	Stent	USDA	PI 476678	MO-6635032	XXXX	TS
Sorghinae	*Dichanthium*	*sericeum*	(R. Br.) A. Camus	USDA	PI 213880	MO-6635033	XXXX	TS
Arundinelleae	*Arundinella*	*hirta*	(Thunb.) Tanaka	USDA	PI 246756	MO-6647156	XXXX	W
Arundinelleae	*Arundinella*	*hookeri*	Munro ex Keng	EK	Kew #0050290	N/A	XXXX	W
Paspalinae	*Paspalum*	*vaginatum*	Sw.	USDA	PI 509022	MO-6635035	XXXX	W
Otachyriinae	*Steinchisma*	*decipiens*	(Nees ex Trin.) W. V. Br.	USDA	PI 462236	MO-6635036	XXXX	TS
Arthropogoinae	*Coleataenia*	*prionitis*	(Nees) Soreng	USDA	PI 496395	MO-6635037	XXXX	W
Tristachyideae	*Danthoniopsis*	*dinteri*	(Pilg.) C.E. Hubb.	USDA	PI 207548	MO-6635038	XXXX	W
	Eriachne	*aristidea*	F. Muell.	USDA	PI 238306	MO-6635039	XXXX	W
	Aristida	*congesta*	Roem. & Schult.	USDA	PI 364389	MO-6635040	XXXX	W
	Aristida	*purpurea*	Nutt.	USDA	PI 598972	N/A	XXXX	W

Source abreviations: MSB=Millenial Seed Bank, NS=Nindethana Australian Seeds, KD=K.M. Devos, DL=D.B. Lowery, LEB=L.E. Bartley, AS=A.J. Studer, ML= M. Ludwig, JRB=J.R. Burkhalter, MW=W.M. Whitten, EK=E.A. Kellogg. Library methods: TS=TruSeq Stranded mRNA, W=Wang, et al. (2011).

Supplemental Table S3.2 - Total orthologs found on each *Sorghum bicolor* and *Setaria italica* chromosome separated by matrix occupancy and orthology inference method.

		Original	Genome-guided			Agalma			Yang & Smith 1 to 1			Yang & Smith MO		
			8_spp	90%	Full	8_spp	90%	Full	8_spp	90%	Full	8_spp	90%	Full
		Genes	Genes	Genes	Genes	Genes	Genes	Genes	Genes	Genes	Genes	Genes	Genes	Genes
S.bicolor Chromosomes	1	3,289	2,018	511	103	1,240	398	106	871	368	167	1,399	375	198
	2	2,181	1,249	293	77	749	257	80	511	201	108	817	210	127
	3	2,587	1,498	335	61	885	275	67	583	235	114	979	241	135
	4	2,112	1,264	297	61	778	250	61	491	198	89	823	193	106
	5	681	294	50	6	183	43	7	103	39	14	200	40	17
	6	1,558	876	177	32	558	166	50	361	128	68	604	139	85
	7	1,073	583	132	19	361	117	22	210	96	45	365	98	53
	8	753	375	78	12	239	76	15	125	48	23	234	58	30
	9	1,434	828	160	25	461	136	30	314	127	58	476	113	69
	10	1,365	751	174	38	456	118	33	284	110	54	470	115	66
Total*		17,033	9,736	2,207	434	5,910	1,836	471	3,853	1,550	740	6,367	1,582	886
S. italica Scaffolds	1	2,112	1,264	297	61	778	250	61	491	198	89	823	193	106
	2	2,172	1,246	294	76	748	255	80	505	198	107	813	209	127
	3	2,091	1,191	246	39	710	219	58	474	192	95	744	188	115
	4	1,365	751	174	38	456	118	33	284	110	54	470	115	66
	5	2,587	1,498	335	61	885	275	67	583	235	114	979	241	135
	6	1,073	583	132	19	361	117	22	210	96	45	365	98	53
	7	1,633	881	172	30	541	156	37	323	113	53	559	121	66
	8	690	295	47	6	186	45	7	103	37	15	204	40	19
	9	3,298	2,023	512	103	1,244	398	106	872	368	167	1,400	375	198
Total*		17,021	9,732	2,209	433	5,909	1,833	471	3,845	1,547	739	6,357	1,580	885

* Differences in totals due to incomplete genome assemblies

Supplemental Table S3.3 - Total orthologs found in each method for the Grape data set with at least four species as the cutoff.

Genome-guided				Agalma				Yang & Smith 1 to 1				Yang & Smith MO			
Genes		Amino Acids		Genes		Amino Acids		Genes		Amino Acids		Genes		Amino Acids	
Total	Min	Total	Min	Total	Min	Total	Min	Total	Min	Total	Min	Total	Min	Total	Min
1,677	437	642,010	112,568	12,744	5,151	4,608,990	1,686,999	13,342	3,939	3,926,469	1,317,706	17,181	6,796	5,871,359	2,713,463

CHAPTER 4: THE SUB-TYPES OF C_4 PHOTOSYNTHESIS: A TRANSCRIPTOMIC AND EVOLUTIONARY APPROACH TO UNDERSTANDING THEM IN THE GRASSES.

Abstract

C_4 photosynthesis is considered the most productive way plants turn sunlight into chemical energy. However, C_4 exists in multiple sub-types, and which sub-type, or combination of sub-types, is most efficient remains unknown. Variations on C_4 photosynthesis are classically divided into three biochemical sub-types: NADP-ME, PCK, and NAD-ME. Recent literature has suggested that many plants don't use a single sub-type exclusively but a mixture of sub-types together or even perhaps different sub-type mixtures under different environmental conditions. One of the three classical sub-types, PCK, is commonly mixed with the other sub-types, and may even account for as much as 25% of photosynthesis in Corn, which has always been considered exclusively NADP-ME. Some researchers have suggested that PCK is always mixed with other sub-types and should not be considered its own sub-type at all: in fact, this view point which has become dominant within the C_4 community in recent years. We conducted a careful investigation of mRNA abundance levels in mesophyll and bundle sheath cells of closely related species from each of the classical C_4 sub-types. We also sampled a close C_3 relative to all of the species for comparison. Our data indicate that while some species clearly mix traditional C_4 sub-types, others show little to no evidence of mixing. Of note, the PCK species we sampled appears to be extremely dominant for the PCK sub-type with very little indication of any of the other sub-types. We conclude that the PCK sub-type is likely functioning on its own and should not be excluded from C_4 sub-type classification systems. Comparative phylogenetic analyses within our Paniceae species also indicate that the most recent common ancestor of the species here sampled likely contained the functional building blocks of all three C_4 pathways

Introduction

The C_4 photosynthetic pathway was first understood biochemically in the 1960's through the discoveries of researchers around the globe; most notably Hatch and Slack (1966). Detailed descriptions of the history of this discovery can be found in Hatch (1992), Furbank (2016), von Caemmerer, et al. (2017), and others. Soon after the biochemical elucidation of the C_4 pathway, it became apparent that not all C_4 species used the same biochemistry. Three distinct biochemical pathways associated with different C_4 species were described and named as NADP-ME, PCK, and NAD-ME, after their respective decarboxylation enzymes (Edwards, et al. 1971, Hatch, et al. 1975, Hatch and Kagawa 1976, Furbank 2016). There was some early discussion about whether the sub-types were mutually exclusive or if one species might employ two or more sub-types together, but in general, the sub-types have been thought of and described in the literature as more or less non-overlapping (Furbank 2016). This view may have been due to the necessity for oversimplification in describing and studying these pathways, as well as experimental evidence that some species, particularly in the grasses, appear to be extremely dominant if not exclusive for one sub-type or another (Gutierrez, et al. 1974, Prendergast, et al. 1987, Lin, et al. 1993). Whatever the reason, for several decades this description of three sub-types has been the standard treatment for C_4 in the literature, and are even used in taxonomic classification and description (Brown 1977).

In more recent years, evidence for wide spread C_4 sub-type mixing has accumulated. Corn (*Zea maize*), for example, has usually been thought of as exclusively NADP-ME, but recent studies have shown that components of the PCK sub-type are present in corn and may be responsible for 10-25% or more of its photosynthetic activity

(Walker, et al. 1997, Wingler, et al. 1999, Majeran, et al. 2010, Furbank 2011, Pick, et al. 2011, Wang, et al. 2014, Koteyeva, et al. 2015, Weissmann, et al. 2016). It also appears that mixing sub-types may allow for more efficient photosynthesis. One explanation for this efficiency is that the use of multiple transport molecules (those from two of the C_4 sub-types rather than just those from one) decreases concentration gradients and hence the energy needed for transport of metabolites between mesophyll (MS) and bundle sheath (BS) cells (Wang, et al. 2014). The potential benefits of sub-type mixing, and even plasticity, with the use of one or more subtypes depending on environmental conditions, remain largely unexplored. As noted by Robert Furbank, "Which [C_4 sub-type or mixture of sub-types] is the 'best' or most efficient way of carrying out C_4 photosynthesis is unknown, and a better understanding of this seems pivotal for future crop engineering strategies" (Furbank 2016, page 4061).

Some of the evidence for sub-type mixing has also led to the observation that many plants with the PCK sub-type also use one of the other two sub-types as well to perform photosynthesis (Furbank 2011). Recently, the suggestion has in fact been made that PCK may never function as a distinct sub-type but instead as an important accessory pathway to NAD-ME or NADP-ME in order to boost photosynthetic efficiency (Furbank 2011, Bräutigam, et al. 2014, Wang, et al. 2014, von Caemmerer and Furbank 2016). Most of the evidence for this idea is based on computational modeling of what we currently know about the C_4 pathways, and as it happens, what we know comes mainly from studying species that are traditionally classified as NADP-ME or NAD-ME, with relatively little biochemical work done on traditionally PCK dominant species (Washburn, et al. 2015). However, experimental evidence from the older literature

supports certain species being at least extremely PCK dominant, and these experiments have yet to be repeated or disproven (Gutierrez, et al. 1974, Prendergast, et al. 1987, Lin, et al. 1993). Several alternative options have been suggested for sub-type classification (Wang, et al. 2014, Washburn, et al. 2015, Rao and Dixon 2016). First, since we know there are over 60 independent origins of C_4 photosynthesis, one could simply assume that each of these origins is distinct enough from the others to be considered its own sub-type. At some level this view is probably the most correct, but also the least useful. A second option is to use a two sub-type system in which NADP-ME and NAD-ME are the only sub-type groups considered and anything currently classified as PCK would be placed into one of these two (Wang, et al. 2014). A third option is a four sub-type classification with NADP-ME, NAD-ME, NADP-ME + PCK, and NAD-ME + PCK as the sub-types (Rao and Dixon 2016). At present, none of these classification systems has been fully adopted by the community, and each researcher uses the system they see as most useful. Part of the reason for this lack of agreement within the community may be that previous studies comparing the classical NADP-ME, NAD-ME, and PCK sub-types have often used representative species for the three sub-types that are not closely-related evolutionary. While these studies have produced interesting and useful results, the choice of species sampling precludes more evolutionarily informed analyses that could be done if closely related species from the three sub-types were used.

To better understand the extent of C_4 sub-type mixing, the role of the PCK pathway in C_4 photosynthesis, and the utility of the different sub-type classification systems, we performed a careful mRNA expression analysis on BS and MS enriched samples across phylogenetically-spaced C_4 plants that are traditionally defined as using

one of each of the C_4 sub-types exclusively, or nearly so. These analyses were performed within the grass tribe Paniceae (Poaceae), the only known group of organism containing all three C_4 sub-types as more closely related to each other (as a monophyletic group) than they are to any C_3 species (Sage, et al. 2011).

Materials and Methods

Plant Materials

Accessions of five plant species were used in this study: *Setaria italica* yugu1, *Urochloa fusca* LBJWC-52, *Panicum hallii* FIL2, *Digitaria californica* PI 364670, and *Sacciolepis indica* PI 338609. More details on each of the accessions, their sources, voucher specimens, and other information can be found in Washburn, et al. (2015) with exception of *P. hallii* FIL2 which was not a part of that study but was obtained from Thomas Juenger of the University of Texas at Austin. Further details about this accession can be found under *Panicum hallii* v2.0, DOE-JGI, http://phytozome.jgi.doe.gov/.

All plant materials were grown in controlled growth chambers at the University of Missouri in Columbia. Plants were grown under 16 hours of light (from 6:00-20:00) and 8 hours of darkness with temperatures of 23C during the day and 20C at night. Lights were placed between 86-88 cm above the plants. Plantings were grown in 4 replicates in a completely randomized design with 32 plants per replicate (except for the case of *Sacciolepis indica* where the plants were smaller and grown with 64 plants per replicate). The third leaf of the plant was sampled between 11:00 and 15:00; we then employed established leaf rolling and mechanical BS isolation methods with some modifications as

described in the Supplemental Material (Sheen and Bogorad 1985, Chang, et al. 2012, Covshoff, et al. 2013, John, et al. 2014) in preparation for RNA extraction.

Sequencing

RNA was extracted using the PureLink® RNA Mini Kit (Invitrogen, Carlsbad, CA, USA) and mRNA-seq libraries were constructed and sequenced by the University of Missouri DNA Core Facility using the TruSeq Stranded mRNA Sample Prep Kit (Illumina, Inc., San Diego, CA, USA) and the Illumina HiSeq and NextSeq platforms.

Analysis

Each mRNA sample was quality trimmed and mapped to the *Sorghum bicolor* genome (Paterson, et al. 2009, DOE-JGI 2017) using Trimmomatic and Trinity following the workflows outlined on their website (Grabherr, et al. 2011, Haas, et al. 2013, Bolger, et al. 2014). This processing included the use of RSEM and Bowtie2 for read mapping and counting as well as edgeR and DESeq for differential expression analysis (Robinson, et al. 2010, Li and Dewey 2011, Langmead and Salzberg 2012, McCarthy, et al. 2012, Love, et al. 2014). A list of known C_4 photosynthesis genes was compiled based on the literature; a custom script and BLAST were then used to find the appropriate homologous genes for each species in order to compare their relative abundance levels (Camacho, et al. 2009, Chang, et al. 2012, Covshoff, et al. 2013, Bräutigam, et al. 2014, John, et al. 2014, Tausta, et al. 2014, Rao, et al. 2016). Ancestral state reconstructions where

performed using the "Trace Character History" command in Mesquite (Maddison and Maddison 2017).

Transcript Normalization and Transcriptome Size Estimation

Relative transcript abundance comparisons within a single sequenced library require only a simple standard normalization (Coate and Doyle 2010). TPM (Transcripts Per Kilobase Million) values as generated using the Trinity and RSEM software packages were here used for all within library comparisons. On the other hand, comparisons between differing cell types or species require further normalization for transcriptome size (Coate and Doyle 2010, Coate and Doyle 2015). Different approaches for this normalization have been suggested, each with benefits and drawbacks.

For comparisons across all cell types and species within our study, we used the Trimmed Mean of M-values (TMM) method described by Robinson and Oshlack (2010) as implemented in DESeq. This method is entirely computational and requires no biological knowledge or additional experiments to implement. However, it relies on the assumption that most genes are not differentially expressed, which is likely not true in many situations (Coate and Doyle 2015).

Another approach is to normalize all transcripts to "housekeeping" genes under the assumption that these genes will be expressed similarly in all tissues. This assumption is sometimes violated making the method undesirable for many applications (Nicot, et al. 2005, Coate and Doyle 2015). However, when exploring the relative transcript levels of C_4 genes within BS cells, the biology of the system makes normalization to certain transcripts extremely useful. Rubisco activase (RBCSACT), for example, is an ideal

candidate for normalization because its relative abundance should be highly correlated with turnovers in the Calvin Cycle. Based on this reasoning, we used RBCSACT for normalization in comparisons between the BS transcriptomes from the different species. Our BS cell analyses where done separately using both normalization to RBCSACT as well as TMM and the results were qualitatively identical.

Results and Discussion

Methods Validation

Since this experiment was the first to apply leaf rolling and mechanical BS isolation to several of our species (particularly the C_3 species *S. indica*), it was unclear how successful the procedures would be. For all five of the species, microscopic examination of the mechanically separated cells revealed high levels of purity for bundles sheath cells with very few, if any, other intact cells being found in the preparations (Figure 1). Examination of the rolled leaves used for isolating MS contents also indicated a high level of purity for most samples (Figure 1). However, *S. indica* performed so poorly in the leaf rolling procedure that it could not be used in the study. Despite many attempts, we were unable to roll *S. indica* leaves with sufficient pressure to burst MS cells without mutilating the leaves. It may still be possible to use leaf rolling with *S. indica* and other C_3 species, but it will likely require a more sophisticated rolling system where pressures can be applied exactly, and/or rolling the leaves at a different developmental stage (Furbank, et al. 1985, Leegood 1985). For this study, we substituted

whole leaf RNA-seq data for *S.indica*, from the same growth conditions, for MS enriched RNA-seq data.

In addition to evaluating the purity of our samples under the microscope, we also performed differential expression analyses for known C_4 genes, and looked for expression patterns consistent with the literature (Covshoff, et al. 2013, John, et al. 2014). These analyses showed clear differences in expression levels between MS and BS samples which are consistent with previous studies and our current understanding of C_4 photosynthesis (Figure 2)(Covshoff, et al. 2013, John, et al. 2014). Each of the C_4 species displays a transcriptional profile consistent with performing C_4 photosynthesis while the C_3 species, as expected, shows a profile consistent with C_3 photosynthesis. These expected results serve as further confirmation that the performance of these isolation procedures in our species is consistent with previous experiments (Covshoff, et al. 2013, John, et al. 2014).

Some Species Mix Sub-Types but Others Clearly Do Not

Setaria italica – S. italica is a member of the Cenchrinae subtribe within the Paniceae. This subtribe is classically defined as using the NADP-ME sub-type of C4 (Gutierrez, et al. 1974, Prendergast, et al. 1987, Lin, et al. 1993). Two enzymes that are particularly indicative of this sub-type are NADP-ME within the BS cell type, and NADP-MDH within the MS cell type. Based on previous literature and the classical definition of the C4 sub-types, we would expect transcript abundance levels of these two

enzymes in their respective cell types to be high, and levels of the PCK, NAD-ME, NAD-MDH, ASP-AT, and ALA-AT within the BS to be low.

Figure 3a shows the levels of each of these transcripts within BS and MS cells as well as a simplified pathway diagram of NADP-ME photosynthesis. As expected from the literature, transcript abundance levels for NADP-ME and NADP-MDH within BS and MS cells respectively are high. Transcript abundance levels for PCK, NAD-ME, NAD-MDH, ASP-AT, and ALA-AT within the BS are also low as expected. As a control, RBSCACT has BS transcript abundance levels many times higher than those of the other enzymes (aside from NADP-ME). High RBSCACT levels are expected in the BS cells of all C4 plants. Within the MS cells of S. italica, NADP-MDH is highly abundant as expected for a plant using the NADP-ME sub-type. PPDK, PEPC, and CA also have moderate to high transcript abundance levels within the MS cells as expected in all C4 plants. Transcript abundance levels for ASP-AT, and ALA-AT are also present at about half the level to NADP-MDH. Abundance of this magnitude is not necessarily expected in an NADP-ME species, but without higher levels of PCK or NAD-ME in the BS, these levels are not likely on their own to indicate the use of other sub-types.

Our data are in agreement with the literature that S.italica is likely using the NADP-ME pathway exclusively or nearly so. Given our data, S.italica could be easily and informatively classified using any of proposed classification systems (discussed in the introduction) or the traditional three-subtype system.

Urochola fusca – U. fusca is a member of the subtribe Melinidinae and is classically defined as using the PCK sub-type (Gutierrez, et al. 1974, Prendergast, et al. 1987, Lin, et al. 1993). As discussed earlier and contrary to the older literature, several

authors and at least one modeling study have suggested that the PCK sub-type is unlikely to function alone in any species (Furbank 2011, Bräutigam, et al. 2014, Wang, et al. 2014). Because of these disagreements in the literature, there are two potential expectations for what our data should look like. If PCK is functioning as its own sub-type with little to no help from other sub-type pathways, one would expect to see high transcript abundance levels for PCK, ASP-AT, and ALA-AT, and low abundance levels for NADP-ME, NAD-ME, and NAD-MDH within BS cells. One would also expect high levels of ASP-AT and ALA-AT within MS cells. Conversely, if PCK is simply ancillary to one of the other sub-types, one would expect to see high levels of NAD-ME and NAD-MDH, or high levels of NADP-ME and NADP-MDH. Our data support the first scenario with high levels of PCK and low levels of NADP-ME and NAD-ME in BS cells (Figure 3b). We also see low to moderate levels of ASP-AT and ALA-AT in BS cells and high levels in MS cells. NAD-MDH levels are also very low in BS cells further supporting the idea that the NAD-ME sub-type is not operating at a high level within U. fusca. Strikingly, these results are very similar to results generated decades ago within the Melinidinae generated using enzyme activity measurements (Gutierrez, et al. 1974, Prendergast, et al. 1987, Lin, et al. 1993). Together, our data strongly suggest that PCK is in fact functioning as the primary C4 sub-type in U.fusca. The relatively low transcript abundance levels of NADP-ME, NAD-ME, and NAD-MDH within U.fusca bundles sheath cells and NADP-MDH in MS cells cannot completely rule out the possibility of these pathways contributing to C4 function in this species, but they do suggest that any contribution is probably many times lower than that of PCK.

These data for U.fusca only fit well within the classical definition and the use of a PCK sub-type. A two or four sub-type system as described above is unable to confidently place U.fusca in any category, because the next highest transcript abundance level after PCK is many times smaller, and NADP-ME and NAD-ME display transcript abundance levels that are not statistically different.

Panicum hallii – The third species we assayed was P. hallii. It is a member of the Panicinae subtribe and is classically defined as using the NAD-ME sub-type. Our expectations for transcript abundance levels within an NAD-ME species are as follows. Within the BS, we expect to see high levels of the NAD-ME, NAD-MDH, ASP-AT, and ALA-AT transcripts. We also expect to see low levels of both PCK, and NADP-ME transcripts within BS cells. For the MS cells of an NAD-ME species we expect to see high levels of ASP-AT and ALA-AT, and low levels of NADP-MDH.

From our data, it appears that some of these expectations are met while others are not (Figure 3c). Transcript abundance levels in both MS and BS cells of P.hallii show strong signs of the NAD-ME sub-type being functional (high levels of NAD-ME, NAD-MDH, ASP-AT, and ALA-AT as described above). Surprisingly though, high levels of NADP-ME and moderate levels of PCK are also seen in the P.hallii BS samples. This combination is unexpected given both the current dogma surrounding the NAD-ME sub-type and the classical literature (Gutierrez, et al. 1974, Prendergast, et al. 1987, Lin, et al. 1993). Interestingly, similar results have been found in recent work on switchgrass (Panicum Virgatum), a close relative to P. hallii and another member of the subtribe Panicinae (Zhang, et al. 2013, Meyer, et al. 2014, Rao and Dixon 2016, Rao, et al. 2016). Rao, et al. (2016) suggest that these high levels of NADP-ME may not contribute

functionally to photosynthesis in Switchgrass, based on the homology of this particular NADP-ME gene to one of the NADP-ME genes in maize that has been shown to be non-photosynthetic. The transcript levels of this NADP-ME isoform in Switchgrass leaves are also much lower in comparison to NAD-ME in BS cells. Rao, et al. (2016) speculate that post-transcriptional or translational modification may account for this. Conversely, in both our study and the Rao study, the NADP-ME transcripts appear to be from the same isoform as that used for photosynthesis in S.itilica or Setaria viridis respectively, an argument for their functional relevance. Further work will be needed to determine if P.hallii is functionally mixing sub-types or if it actually falls into the classical NAD-ME definition functionally.

Digitaria californica – D. californica is a member of the Anthephorinae subtribe in the Paniceae. This subtribe is classified as using the NADP-ME sub-type. Our expectations for transcript abundance levels within this group are the same as those for the S.italica comparisons discussed earlier. We expect D. californica to display high levels of NADP-ME and NADP-MDH within the BS and MS specific samples respectively. We also expect to see low levels of PCK, NAD-ME and NAD-MDH within the BS samples and low levels of ASP-AT and ALA-AT within both MS and BS samples. Figure 3d shows our results for D. californica. As expected, levels of NADP-ME and NADP-MDH are high in BS and MS samples respectively, suggesting that D.californica utilizes the NADP-ME sub-type. However, D. californica displays unexpectedly high transcript abundance levels for transcripts associated with the PCK sub-type. In this case, PCK, ASP-AT, and ALA-AT all show high levels of transcript abundance, and for the most part these high levels are found in the expected locations

(MS or BS) for a functional PCK pathway. The exception to these expectations is in the BS cells, where ASP-AT and ALA-AT are low, as one might anticipate for an NADP-ME sub-type species. PCK transcript abundance levels are more than double those of NADP-ME in the BS cells, and ASP-AT levels in the MS cells are also extremely high. NAD-ME and NAD-MDH levels are barely detectable in the BS cells of D. californica, suggesting the NAD-ME sub-type is unlikely to function at a meaningful level within this species. As in the case of P.hallii, further work is needed to determine if D.californica is mixing the NADP-ME and PCK sub-types, but our data support this hypothesis.

Sacciolepis indica – The final species sampled was S. indica, a close C3 relative to all four of the C4 species considered above. Since this species uses C3 photosynthesis, we expect to see very low levels of all C4-related transcripts. Although our data were generally consistent with this expectations, the levels were not as low as one might have expected. In many cases, they were higher than the base levels seen for the same transcripts in some of the C4 species examined. For example, PCK and NAD-ME levels in S. indica BS cells were higher than they were in S. italica BS cells. These low levels of C4 transcript abundance and their occurrence in the correct places for C4 photosynthesis suggest the possibility that these biochemical pathways might be operating at some low level in S. indica. Our data also suggest that there is no strong preference for one of the sub-type pathways over the others within S. indica (Figure 4).

Ancestral State Reconstruction for the MPCA Clade

How each of the different sub-types (or mixes of sub-types) within the MPC(A) clade evolved is an intriguing question, with implications for crop improvement and engineering. Hypotheses about this clade's evolution have been put forward in the past (Figure 5), but testing them has been challenging (Washburn, et al. 2015). Our data provide a new opportunity for examining the evolution of C_4 sub-types within the MPC(A) clade in the light of transcript abundance levels from different cell types and representatives of each of the three C_4 sub-types and a close C_3 relative. Ancestral state reconstructions of transcript abundance levels at each node of the MPC(A) phylogeny show mixtures of all three primary sub-type enzymes at every node (Figure 6). The appearance of transcriptional level sub-type mixing in at least two of our four species also supports the common ancestor of these species having employed sub-type mixing. Additionally, transcripts associated with multiple enzymes involved in each of the three C_4 sub-types are present at low, but higher than expected, levels (given their low levels in some of the C_4 species here sampled) within the C_3 species *S. indica*. Taken together, these data support the hypothesis that the C_4 sub-types existed together within the MRCA of the MPC(A) clade of the Paniceae, rather than each having evolved independently from the others or one having evolved first with the other two evolving from it in a stepwise fashion.

Conclusions

A phylogenetically-aware analysis of BS and MS transcript abundances across closely related C4 species representing each of the classical C_4 sub-types and a close C_3

relative were performed. The results indicate that: 1) MPC(A) representatives of each of the classical C_4 sub-types are transcriptionally distinct from each other. 2) *S.italica* and *U.fusca* have transcript abundance levels consistent with the classically defined NADP-ME and PCK sub-types respectively with little to no sub-type mixing. 3) Since *U.fusca* appears to be using the PCK sub-type at a nearly exclusive level, it follows that the traditional PCK sub-type classification is, in fact, biologically relevant and should be considered in studies of C_4 photosynthesis. 4) *P. hallii* and *D.californica* show transcript abundance levels that are indicative of sub-type mixing within these species and further examination on a protein level should be conducted to confirm this occurrence. 5) *S.indica* (a C_3 species) has BS transcript abundance levels consistent with pre-adaptation to C_4 photosynthesis, potentially for all three sub-types (Gould 1989, Christin, et al. 2009, Christin, et al. 2015, Washburn, et al. 2016). 6) Ancestral state reconstructions and the *S.indica* transcript abundance levels are most consistent with the MRCA of the MPC(A) clade using all three sub-types at some level: these analyses do not support an independent origin of the sub-types in the phylogeny nor step-wise evolution of one sub-type from another.

While a great deal of effort has gone into understanding and engineering C_4 photosynthesis into C_3 species and in improving it in species that already have it, we suggest more research is needed both on the natural diversity C_4 sub-types and on how mixing them together might increase photosynthetic efficiency, drought tolerance, and crop productivity. The PCK sub-type, in particular, remains extremely understudied. Given PCK's importance to photosynthesis in corn, arguably the world's most productive

crop plant, a better understanding of PCK has great potential to aid in crop improvement and engineering.

References

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30.

Bräutigam A, Schliesky S, Külahoglu C, Osborne CP, Weber APM. 2014. Towards an integrative model of C4 photosynthetic subtypes: insights from comparative transcriptome analysis of NAD-ME, NADP-ME, and PEP-CK C4 species. J. Exp. Bot. 65:3579-3593.

Brown WV. 1977. The Kranz syndrome and its subtypes in grass systematics. Memoirs of the Torrey Botanical Club 23:1-97.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden T. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10:421.

Chang YM, Liu WY, Shih AC, Shen MN, Lu CH, Lu MY, Yang HW, Wang TY, Chen SC, Chen SM, Li WH, Ku MS. 2012. Characterizing regulatory and functional differentiation between maize mesophyll and bundle sheath cells by transcriptomic analysis. Plant Physiol. 160:165-177.

Christin P-A, Samaritani E, Petitpierre B, Salamin N, Besnard G. 2009. Evolutionary insights on C4 photosynthetic subtypes in grasses from genomics and phylogenetics. GBE 1:221-230.

Christin PA, Arakaki M, Osborne CP, Edwards EJ. 2015. Genetic enablers underlying the clustered evolutionary origins of C4 photosynthesis in angiosperms. Mol. Biol. Evol. 32:846-858.

Coate JE, Doyle JJ. 2010. Quantifying Whole Transcriptome Size, a Prerequisite for Understanding Transcriptome Evolution Across Species: An Example from a Plant Allopolyploid. GBE 2:534-546.

Coate JE, Doyle JJ. 2015. Variation in transcriptome size: are we getting the message? Chromosoma 124:27-43.

Covshoff S, Furbank RT, Leegood RC, Hibberd JM. 2013. Leaf rolling allows quantification of mRNA abundance in mesophyll cells of sorghum. J. Exp. Bot. 64:807-813.

DOE-JGI. 2017. *Sorghum bicolor* v3.1.

Edwards GE, Kanai R, Black CC. 1971. Phosphoenolpyruvate carboxykinase in leaves of certain plants which fix CO2 by the C4-dicarboxylic acid cycle of photosynthesis. Biochem. Biophys. Res. Commun. 45:278-285.

Furbank RT. 2011. Evolution of the C4 photosynthetic mechanism: are there really three C4 acid decarboxylation types? J. Exp. Bot. 62:3103-3108.

Furbank RT. 2016. Walking the C 4 pathway: past, present, and future. J. Exp. Bot. 67:4057-4066.

Furbank RT, Stitt M, Foyer CH. 1985. Intercellular compartmentation of sucrose synthesis in leaves of Zea mays L. Planta 164:172-178.

Gould SJ. 1989. Wonderful life: the Burgess Shale and the nature of history. 1st ed. New York, W.W. Norton.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 29:644-652.

Gutierrez M, Gracen VE, Edwards GE. 1974. Biochemical and cytological relationships in C4 plants. Planta 119:279-300.

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, MacManes MD, Ott M, Orvis J, Pochet N, Strozzi F, Weeks N, Westerman R, William T, Dewey CN, Henschel R, LeDuc RD, Friedman N, Regev A. 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat. Protocols 8:1494-1512.

Hatch M, Kagawa T, Craig S. 1975. Subdivision of C4-Pathway Species Based on Differing C4 Acid Decarboxylating Systems and Ultrastructural Features. Funct. Plant Biol. 2:111-128.

Hatch M, Slack C. 1966. Photosynthesis by sugar-cane leaves. A new carboxylation reaction and the pathway of sugar formation. Biochem. J. 101:103-111.

Hatch MD. 1992. I can't believe my luck. Photosynthesis Res. 33:1-14.

Hatch MD, Kagawa T. 1976. Photosynthetic activities of isolated bundle sheath cells in relation to differing mechanisms of C4 pathway photosynthesis. Archives of Biochemistry and Biophysics 175:39-53.

John CR, Smith-Unna RD, Woodfield H, Covshoff S, Hibberd JM. 2014. Evolutionary Convergence of Cell-Specific Gene Expression in Independent Lineages of C4 Grasses. Plant Physiol. 165:62-75.

Koteyeva NK, Voznesenskaya EV, Edwards GE. 2015. An assessment of the capacity for phosphoenolpyruvate carboxykinase to contribute to C4 photosynthesis. Plant Sci. 235:70-80.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9.

Leegood RC. 1985. The intercellular compartmentation of metabolites in leaves of Zea mays L. Planta 164:163-171.

Li B, Dewey CN. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12.

Lin C, Tai Y, Liu D, Ku M. 1993. Photosynthetic mechanisms of weeds in taiwan. Funct. Plant Biol. 20:757-769.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15.

Maddison WP, Maddison DR. 2017. Mesquite: a modular system for evolutionary analysis.

Majeran W, Friso G, Ponnala L, Connolly B, Huang M, Reidel E, Zhang C, Asakura Y, Bhuiyan NH, Sun Q, Turgeon R, van Wijk KJ. 2010. Structural and Metabolic Transitions of C(4) Leaf Development and Differentiation Defined by Microscopy and Quantitative Proteomics in Maize. Plant Cell 22:3509-3542.

McCarthy DJ, Chen Y, Smyth GK. 2012. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res. 40:4288-4297.

Meyer E, Aspinwall MJ, Lowry DB, Palacio-Mejía JD, Logan TL, Fay PA, Juenger TE. 2014. Integrating transcriptional, metabolomic, and physiological responses to drought stress and recovery in switchgrass (Panicum virgatum L.). BMC Genomics 15:527.

Nicot N, Hausman J-F, Hoffmann L, Evers D. 2005. Housekeeping gene selection for real-time RT-PCR normalization in potato during biotic and abiotic stress. J. Exp. Bot. 56:2907-2914.

Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A, Schmutz J, Spannagl M, Tang H, Wang X, Wicker T, Bharti AK, Chapman J, Feltus FA, Gowik U, Grigoriev IV, Lyons E, Maher CA, Martis M, Narechania A, Otillar RP, Penning BW, Salamov AA, Wang Y, Zhang L, Carpita NC, Freeling M, Gingle AR, Hash CT, Keller B, Klein P, Kresovich S, McCann MC, Ming R, Peterson DG, Mehboob ur R, Ware D, Westhoff P, Mayer KFX, Messing J, Rokhsar DS. 2009. The Sorghum bicolor genome and the diversification of grasses. Nature 457:551-556.

Pick TR, Brautigam A, Schluter U, Denton AK, Colmsee C, Scholz U, Fahnenstich H, Pieruschka R, Rascher U, Sonnewald U, Weber AP. 2011. Systems analysis of a maize leaf developmental gradient redefines the current C4 model and provides candidates for regulation. Plant Cell 23:4208-4220.

Prendergast H, Hattersley P, Stone N. 1987. New structural/biochemical associations in leaf blades of C4 grasses (Poaceae). Funct. Plant Biol. 14:403-420.

Rao X, Dixon RA. 2016. The Differences between NAD-ME and NADP-ME Subtypes of C4 Photosynthesis: More than Decarboxylating Enzymes. Front. Plant Sci. 7.

Rao X, Lu N, Li G, Nakashima J, Tang Y, Dixon RA. 2016. Comparative cell-specific transcriptomics reveals differentiation of C4 photosynthesis pathways in switchgrass and other C4 lineages. J. Exp. Bot. 67:1649-1662.

Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26.

Robinson MD, Oshlack A. 2010. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 11:R25.

Sage RF, Christin PA, Edwards EJ. 2011. The C4 plant lineages of planet earth. J. Exp. Bot. 62:3155-3169.

Sheen J-Y, Bogorad L. 1985. Differential Expression of the Ribulose Bisphosphate Carboxylase Large Subunit Gene in Bundle Sheath and Mesophyll Cells of Developing Maize Leaves Is Influenced by Light. Plant Physiol. 79:1072-1076.

Tausta SL, Li P, Si Y, Gandotra N, Liu P, Sun Q, Brutnell TP, Nelson T. 2014. Developmental dynamics of Kranz cell transcriptional specificity in maize leaf reveals early onset of C4-related processes. J. Exp. Bot.

von Caemmerer S, Furbank RT. 2016. Strategies for improving C4 photosynthesis. Curr. Opin. Plant Biol. 31:125-134.

von Caemmerer S, Ghannoum O, Furbank RT. 2017. C4 photosynthesis: 50 years of discovery and innovation. J. Exp. Bot. 68:97-102.

Walker RP, Acheson RM, Técsi LI, Leegood RC. 1997. Phosphoenolpyruvate Carboxykinase in C4 Plants: Its Role and Regulation. Funct. Plant Biol. 24:459-468.

Wang Y, Bräutigam A, Weber APM, Zhu X-G. 2014. Three distinct biochemical subtypes of C4 photosynthesis? A modelling analysis. J. Exp. Bot. 65:3567-3578.

Washburn JD, Bird KA, Conant GC, Pires JC. 2016. Convergent Evolution and the Origin of Complex Phenotypes in the Age of Systems Biology. Int. J. Plant Sci. 177:305-318.

Washburn JD, Schnable JC, Davidse G, Pires JC. 2015. Phylogeny and photosynthesis of the grass tribe Paniceae. Am. J. Bot. 102:1493-1505.

Weissmann S, Ma F, Furuyama K, Gierse J, Berg H, Shao Y, Taniguchi M, Allen DK, Brutnell TP. 2016. Interactions of C4 Subtype Metabolic Activities and Transport in Maize Are Revealed through the Characterization of DCT2 Mutants. Plant Cell 28:466-484.

Wingler A, Robert PW, Zhi-Hui C, Leegood RC. 1999. Phosphoenolpyruvate Carboxykinase Is Involved in the Decarboxylation of Aspartate in the Bundle Sheath of Maize. Plant Physiol. 120:539-545.

Zhang J-Y, Lee Y-C, Torres-Jerez I, Wang M, Yin Y, Chou W-C, He J, Shen H, Srivastava AC, Pennacchio C, Lindquist E, Grimwood J, Schmutz J, Xu Y, Sharma M, Sharma R, Bartley LE, Ronald PC, Saha MC, Dixon RA, Tang Y, Udvardi MK. 2013. Development of an integrated transcript sequence database and a gene expression atlas for gene discovery and analysis in switchgrass (Panicum virgatum L.). Plant J 74:160-173.

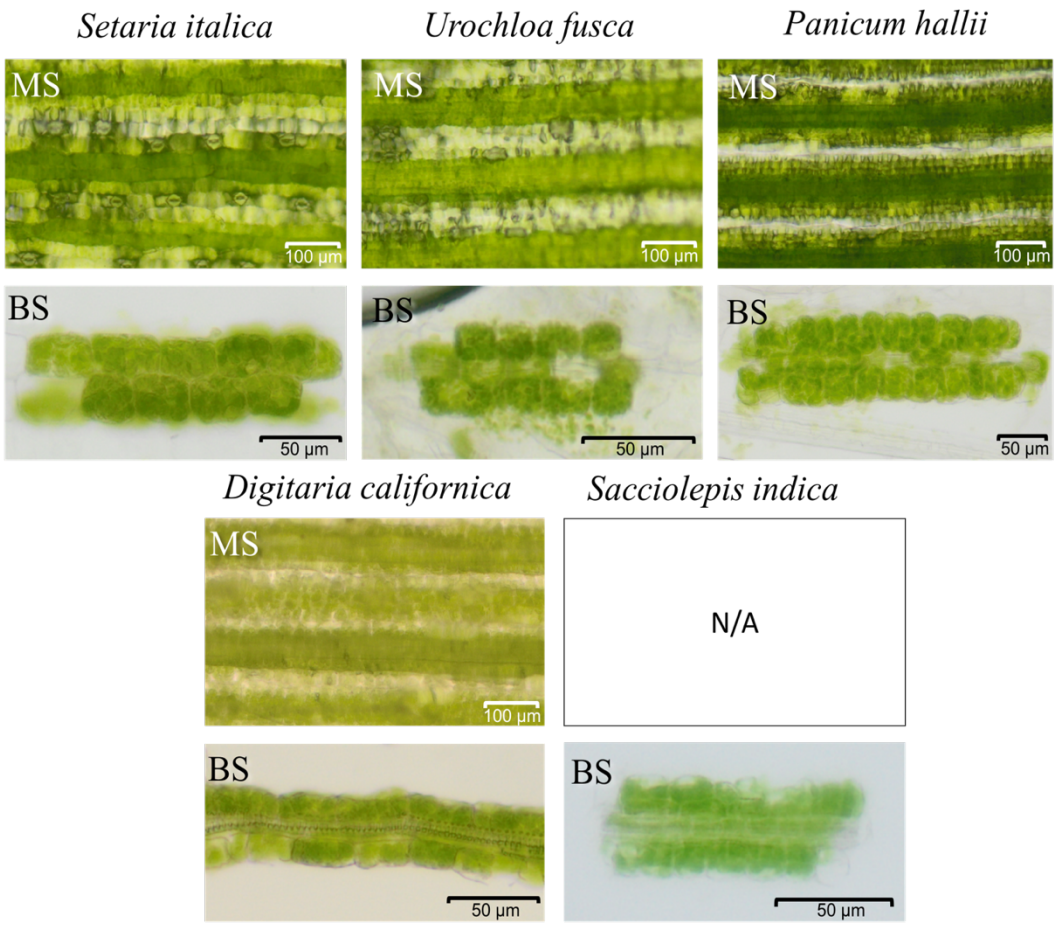


Figure 4.1 - Microscope pictures of mesophyll and bundle sheath preparations.

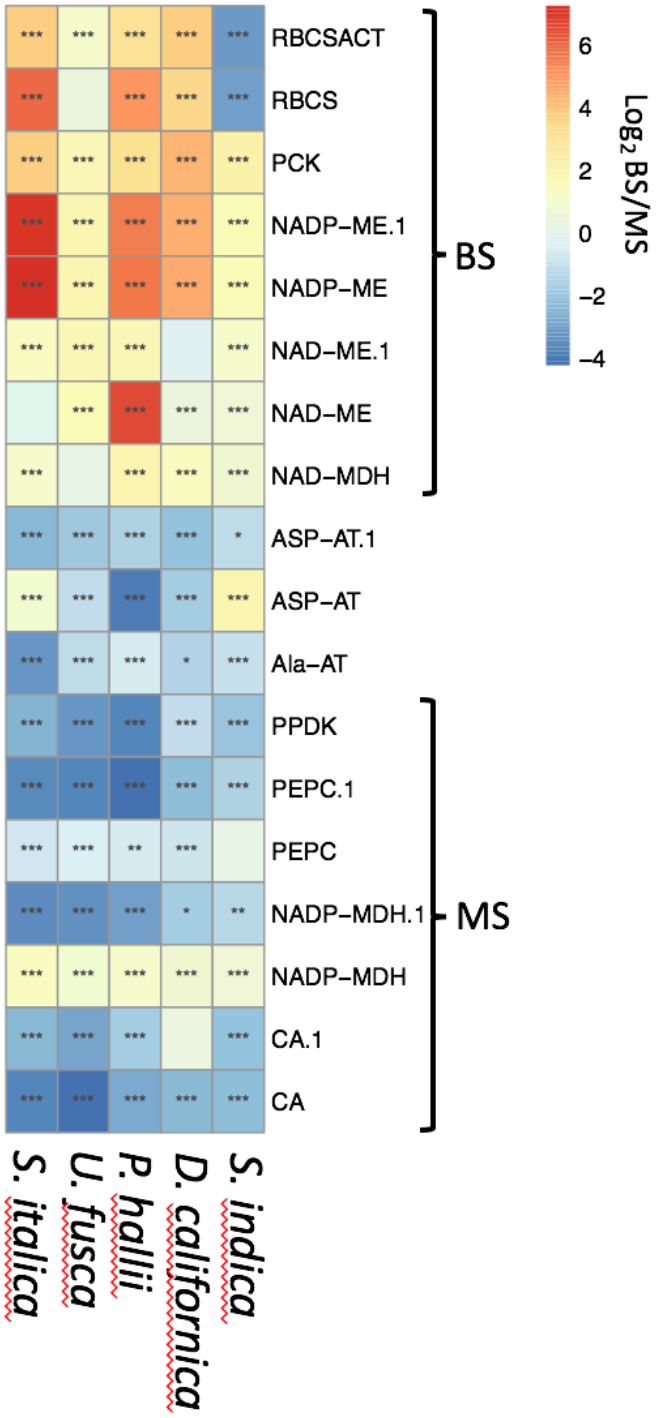


Figure 4.2 - Log2 fold change between mesophyll (MS) and bundle sheath (BS) cells for a subset of well-studied C4 related genes across all five taxa. Brackets indicate genes commonly considered MS or BS specific.

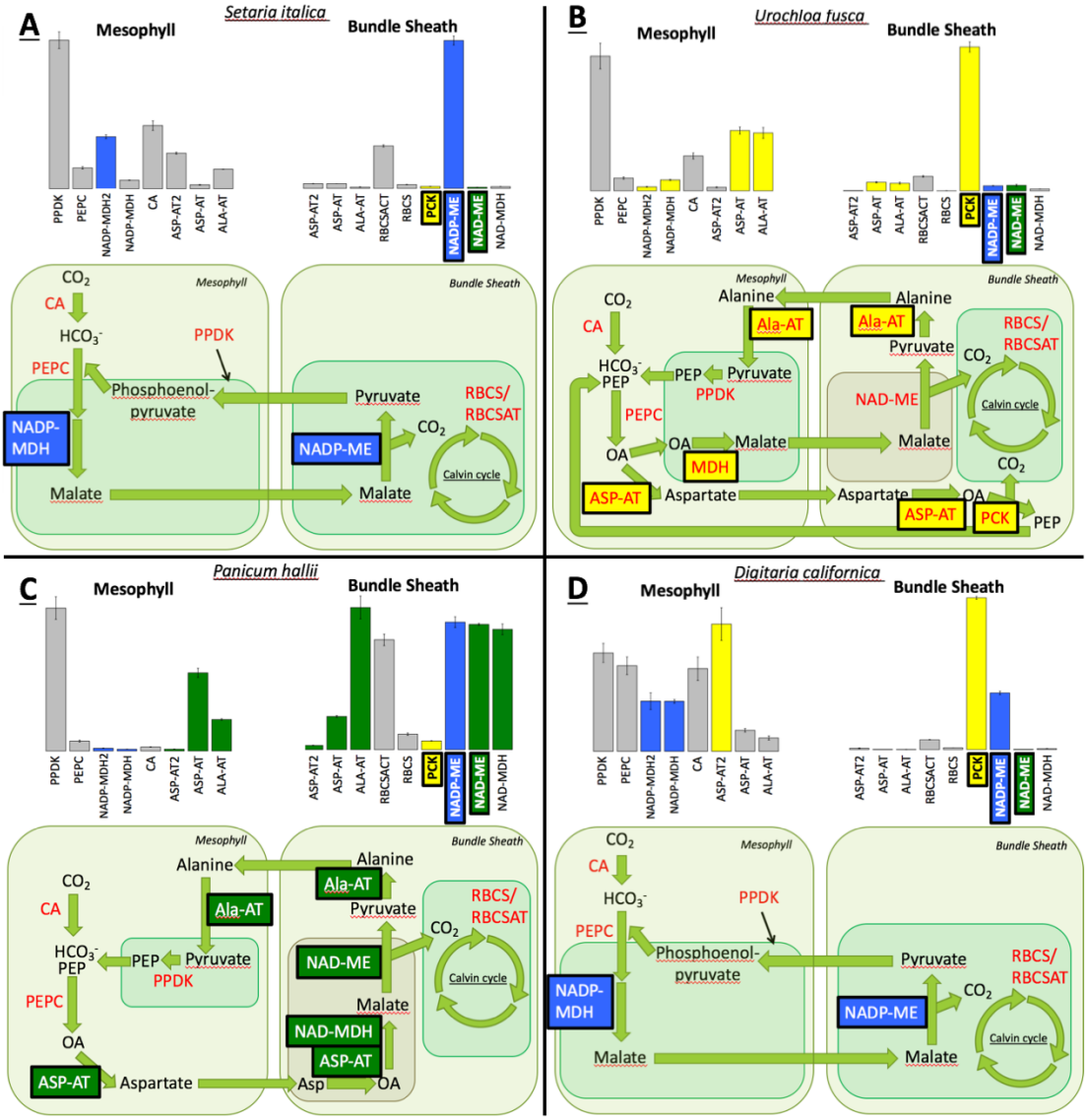


Figure 4.3 - Relative transcript abundance levels for Mesophyll and Bundle Sheath each of C4 species as well as simplified diagrams of the C4 pathways into which each has traditionally been classified.

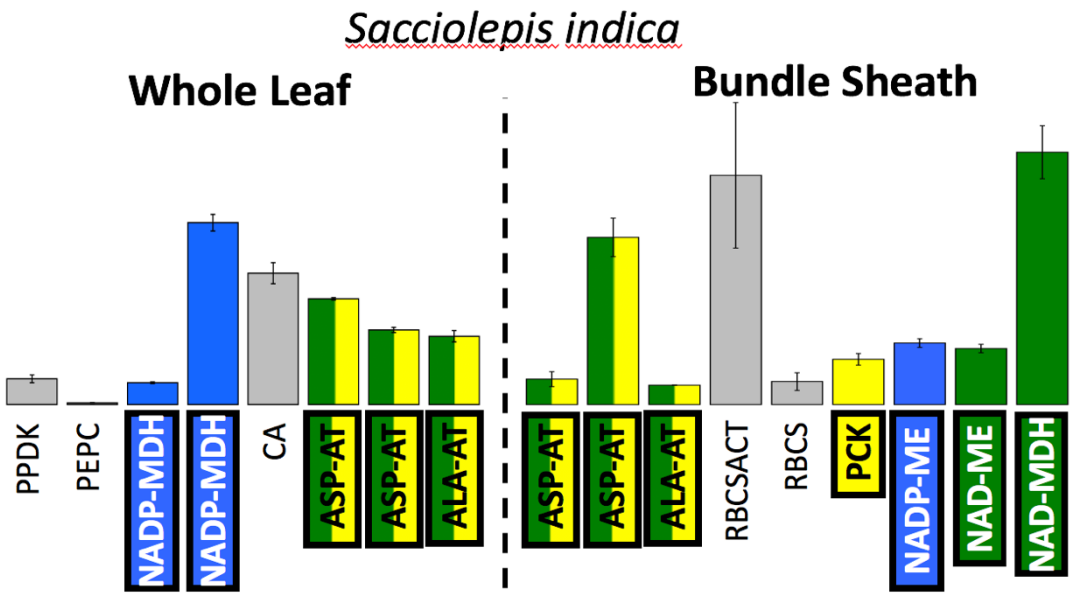


Figure 4.4 - Whole leaf and bundle sheath enriched transcript abundance levels with Sacciolepis indica.

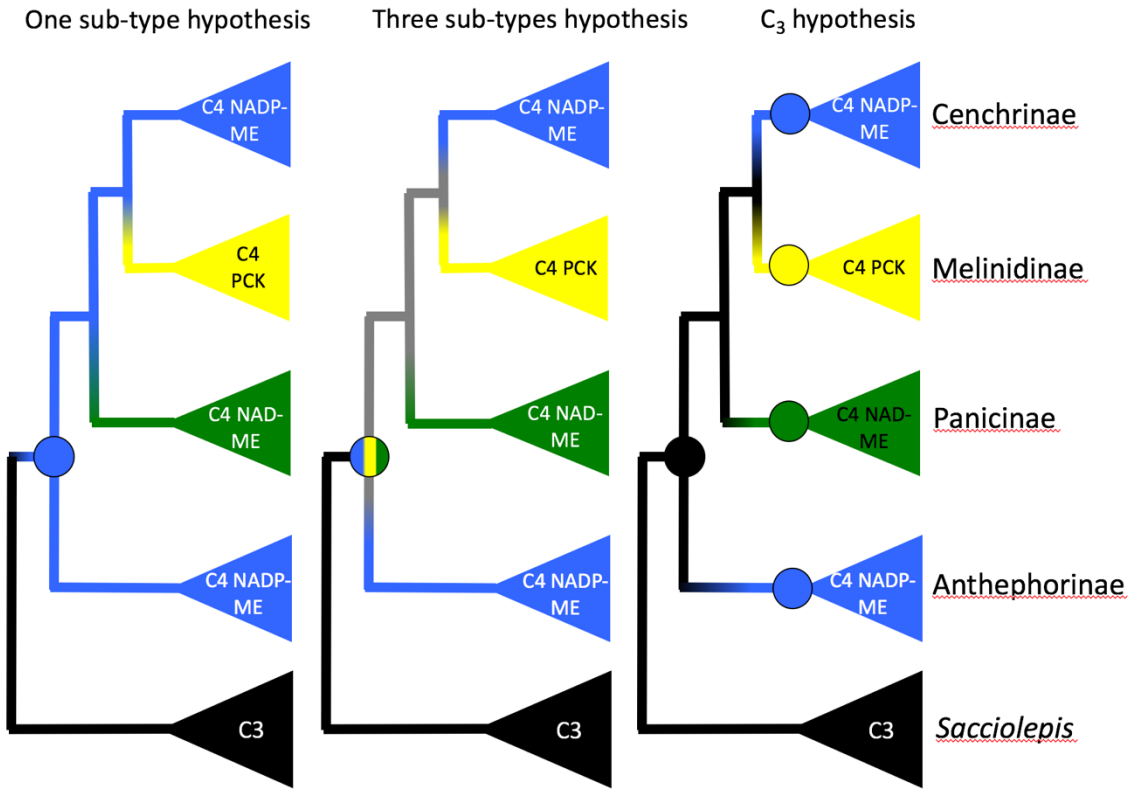


Figure 4.5 - Three hypotheses for the evolution of C4 sub-types within the tribe Paniceae. The one sub-type hypothesis posits that the most recent common ancestor (MRCA) utilized one sub-type exclusively, and the other types evolved from it is a step-wise fashion. The three sub-type hypothesis suggests that all three sub-types existed in the MRCA and then each has become dominant in one clade or another over time. The C_3 hypothesis is based on the idea that each of the sub-types evolved independently from a C_3 ancestor. Figure was modified and re-drawn from Washburn, et al. (2015) and Washburn et al. (in review).

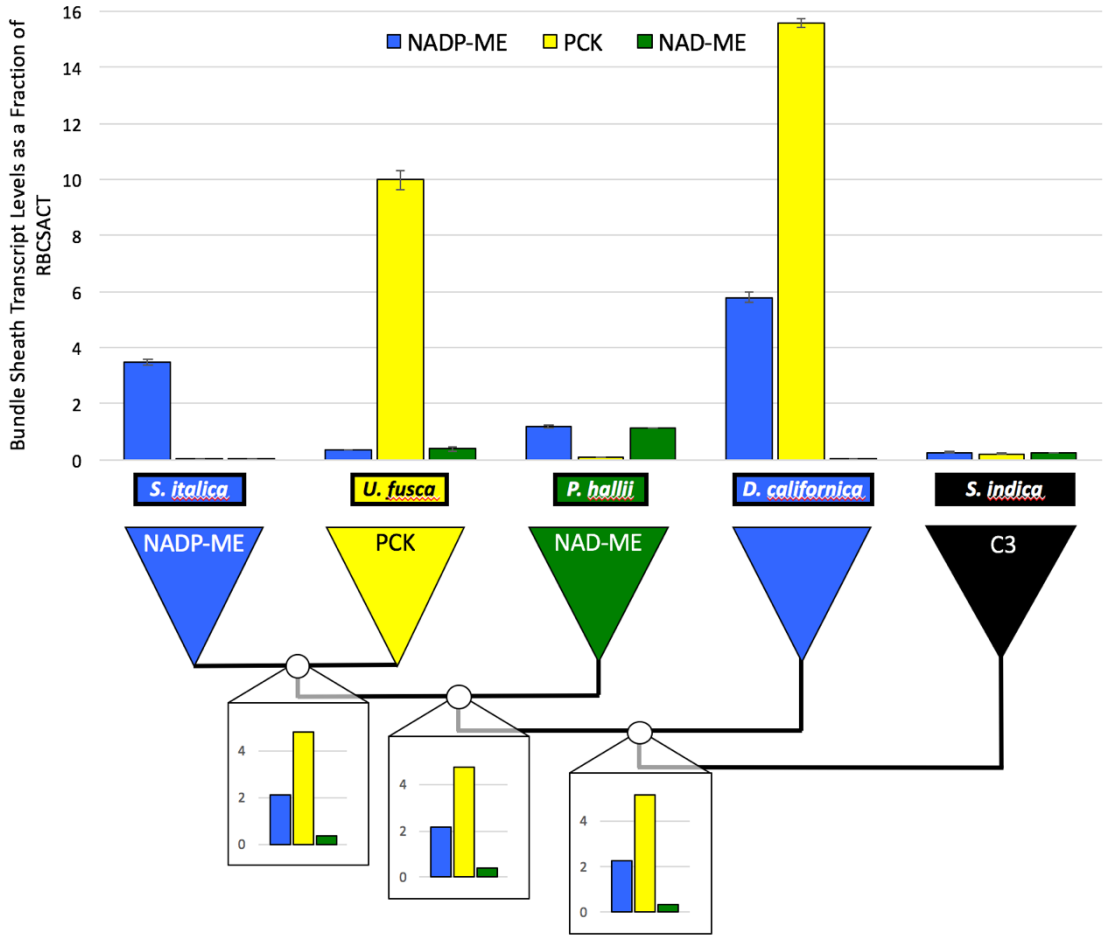


Figure 4.6 - Between species comparisons of the transcript abundances of the NADP-ME, PCK, and NAD-ME transcripts within Bundle Sheath cells along with their nuclear gene phylogenetic relationships and ancestral state reconstructions of transcript abundance levels. Transcript levels are normalized to those of

CHAPTER 5: FUTURE DIRECTIONS

Based on conclusions from the previous chapters there are several avenues of further research which seem pertinent, timely, and most likely to be impactful both scientifically and in crop improvement. I describe three of these research directions which I see as most useful below. They are: 1) Increasing our understanding of the Paniceae phylogeny, 2) Increasing our Understanding of C_4 Sub-Type Evolution, Diversity, and Sub-Type Mixing, and 3) Implications of C4 sub-types for crop improvement.

Increasing our Understanding of the Paniceae Phylogeny

There are two basic strategies for improving our current understanding of the Paniceae phylogeny. The first is greater taxon sampling, and the second is greater depth of gene sampling (meaning whole genome sequencing). The first one is likely to have the greatest impact on our understanding of the phylogeny itself, and could be accomplished by adding more species and/or adding more genotypes of the currently sampled species from the previous chapters. The second option, is unlikely to improve our understanding of the phylogeny directly, but might improve our understanding of the causes of incongruence between gene trees within the tribe.

Adding more species to the current Paniceae sampling from previous chapters is the most likely way to improve confidence in the phylogenetic patterns seen in the tree, but adding more genotypes of species already represented in the tree would potentially improve our ability to differentiate between the causes of gene tree incongruence within

the tribe. Unfortunately, the areas of the tree where greater sampling is most needed are also those for which plant materials are most difficult to obtain. Therefore, such a project would be ideally suited for probe-based approaches, particularly if high enough quality DNA can be extracted from herbarium specimens to include them in the analysis. In my opinion, further chloroplast and mitochondrial gene sampling would likely not be particularly informative for the overall Paniceae phylogeny, although it may be useful for looking more deeply into issues of incongruence between the chloroplast and nuclear trees and for studies within particular subtribes of the Paniceae.

The second option for improving the tribe Paniceae's phylogeny is to add more genes to the current sampling. Given that we already have over 2,000 genes in our phylogeny, the next step would be whole genome sequencing across the clade. This strategy is, in my opinion, unlikely to improve our current understanding of the phylogeny itself, but should be useful for dissecting the causes of incongruence between the gene trees and for understanding which genes are causing the incongruence and what their biological functions are. Such a study would, I think, greatly improve the utility of the Paniceae tree for understanding the evolution of different traits (like C_4 photosynthesis). Adding only a few genomes rather than genomes for the entire tribe, for example one from each of the sub-type clades and *Sacciolepis,* would probably be in itself yield a substantial improvement in out understanding of why the nuclear gene trees are incongruent with each other and with the chloroplast tree. Several genomes needed for this are currently published or in the process of sequencing at the Department of Energy's Joint Genome Institute. Once these genomes are completed and available for use, only an Anthephorinae and *Sacciolepis* genome will be needed to perform a high-

quality assessment of the reasons for incongruence in the tribe, particularly those relating to C4 photosynthesis.

Increasing our Understanding of C_4 Sub-Type Evolution, Diversity, and Sub-Type Mixing

There are several approach's that I think would be useful for further understanding C_4 sub-type evolution. 1) Increased cross-species level RNA-seq sampling within one or more of subtribes Cenchrinae, Melinidinae, Panicinae, and Anthephorinae (MPCA), as well as within the unnamed clade of close C_3 relatives. 2) Population level RNA-seq sampling within one or more C_4 species from within the Paniceae or any other C_4 group.

Increasing controlled and replicated RNA-seq sampling across species in any of the C_4-sub-type clades would enable a much better understanding of that clade's ancestral C_4-sub-type and how variable that sub-type is within the clade. It would also provide clearer evidence of the amount of sub-type mixing going on in that clade. If this sampling were extended to all four of the MPCA subtribes and several of the close C_3 outgroup species it would allow for a much more accurate assessment of the ancestral state of the entire MPCA than we currently have. The inclusion of enzyme activity measurements in those analyses would also go a long way in determining the ancestral state of the MPCA clade. Obtaining these data from whole leaf sampling would probably be sufficient for the simple ancestral state reconstruction and some further analyses, but the use of mesophyll (MS) and bundle sheath (BS) enriched RNA-seq as performed in

Chapter 4 above might provide greater insights. That said, the labor costs of generating the MS and BS enriched data likely outweighs its utility in this situation.

Population level RNA-seq for multiple genotypes of any C_4 species is another sure way of increasing our understanding of C_4 sub-type evolution, diversity, and mixing. In this case, it might also have direct and obvious implications for crop improvement. As discussed in Chapter 4, it has been hypothesized that C_4 sub-type mixing may influence for, and allow further improvement of, C_4 photosynthetic efficiency. The first step in testing this hypothesis, in my opinion, is to obtain a data set with RNA-seq across multiple genotypes of a single species. This might be particularly meaningful if a domesticated species with improved and unimproved cultivars is examined. The simple hypothesis is that the unimproved cultivars will have different ratios of sub-type mixing than the improved ones. Seeing this would indicate that human selection efforts have influenced sub-type mixing within the species. Not seeing this would not mean that sub-type mixing ratios do not influence C_4 photosynthetic efficiency or overall plant fitness, only that it has not been selected for.

Implications of C_4 Sub-Types for Crop Improvement

The idea of C_4 sub-type mixing is, in my opinion, the next logical step that may have impacts on agriculture. As described above, variation for sub-type mixing within a species, and the likelihood of it impacting crop production, should be easily tested with a dataset including a large sampling of RNA-seq from across multiple genotype of one species. Such a data set was recently generated (but is not yet publically available) by the Edward Buckler lab. Examining this data set or others like it is the first step in

determining if manipulation C_4 sub-type mixing might be useful in crop production. Another data set that is currently being generated within the Thomas Juenger lab at UT Austin includes switchgrass genotypes planted across a climactic gradient from the northern to the southern edges of the United States. RNA-seq data generated from this study could also be extremely, telling about the interplay between environment and genotype that control sub-type mixing, and if variation in it effects plant adaptation to specific climates and conditions.

Another direction which I think has potential economic impacts is the introduction of the components of the different sub-type pathways into species not currently using them. For this application, I think that the study of the PCK sub-type has the most potential as it currently the least well understood. Further study of the Melinidinae sub-tribe should be carried out to better understand how the PCK sub-type functions on its own and in conjunction with other sub-types.

APPENDIX 1: ODE TO ENERGY

By Jacob D. Washburn

For 200-plus years, fossil fuels have ruled the day.
They've powered our cars, and the homes where we stay.
They will one day run out, and they cause pollution.
So now we look for a greener solution.

Sustainable energy that won't soon run out,
Is what modern research is all about.
Renewable fuels, they may cost a mint,
But it will be worth it, if we lower our carbon foot-print.

There is solar, and wind, and nuclear too.
Hydro-electric or geothermal would likely do.
But what about plants? They get energy from the sun.
They're cheap, they're simple, and hey, they're kinda' fun!

Which source of power will beat out the rest?
Perhaps a combination of all will be best.
Give it more time, and the answer will be ready.
That is, if we can keep funding rates steady.

The future of technology remains to be seen.
But one thing is likely, no matter the scheme.
To power the lights there are sources not a few.
But to fuel our bodies, only plants will do.

APENDIX 2: C_4 PHOTOSYNTHESIS

By Jacob D. Washburn

Four billion years ago, oxygen on earth was rare.
You and I couldn't breathe if we had been there.
Then along came a molecule that changed earth's face,
RuBisCo, or Ribulose-1,5-bisphosphate carboxylase!

But RuBisCo had a problem that made arid plants blue.
Though it preferred carbon dioxide, it could use oxygen too.
O_2 plus RuBisCo causes photorespiration,
It's a wasteful process, and reduces carbon fixation.

As oxygen levels soared, in hot and dry places,
Photorespiration sent evolution off to the races.
A CO_2 pump was C_4's successful solution,
It appeared 60-plus times in convergent evolution.

Each of C_4's origins has its own composition,
PEPC's the only enzyme found in every rendition.
Anatomical, biochemical, and functional differences abound,
Can we use them in crops to make production systems sound?

C_4 sub-type mixing may make plants better in the field,
Can we fine tune these combos to result in higher yield?
To transformation and to CRISPR this problem seems conducive,
But without biological predictions, it will long remain elusive.

Perhaps your interests are in making food supplies more stable,
Maybe you like evolutionary stories and the discoveries they enable.
Whatever it may be that brought you to this session,
I hope that you learned something, and with that I'll take a question.

VITA

Jacob D. Washburn grew up in the small town of Wales, Utah. At age 19, Jacob served as a missionary for The Church of Jesus Christ of Latter-day Saints; spending two years in Rio Grande do Sul, Brazil. Jacob received an A.S. from Snow College in 2007, a B.S. from Brigham Young University in 2010, and an M.S. in Plant Breeding and Genetics from Texas A&M University in 2012. Upon graduation from the University of Missouri, Jacob, his wife Melinda, and their three children, Nathan, Sam, and Emma, will move to Ithaca, NY where Jacob will begin a postdoctoral position at Cornell University.