The Pennsylvania State University

The Graduate School

# A DISSERTATION IN POACEAE NUCLEAR PHYLOGENY AND EVOLUTION OF C4 PHOTOSYNTHESIS

A Dissertation in

Biology

by

Weichen Huang



Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

May 2023

The dissertation of Weichen Huang was reviewed and approved by the following:

Claude dePamphilis
Professor of Biology
Chair of Committee

Hong Ma
Professor of Biology
Dissertation Advisor

Jesse Lasky
Associate Professor of Biology

Surinder Chopra
Professor of Maize Genetics

Stephen W. Schaeffer
Professor of Biology
Associate Department Head of Graduate Education

# ABSTRACT

Phylogenetics are fundamental to contemporary biology and have offered novel insights into evolutionary questions. Early phylogenetic studies relied on morphological characters to resolve species relationship, while recent studies incorporated molecular data. To date, many branches on the tree of life have been depicted, and the tree is continuously growing as more species are discovered, described and included. Arguably the most famous example is the resolution of the origin of *Homo sapiens* in Africa, supported by numerous fossil records as well as phylogenetic analyses.

In the field of plant biology, crop plants have long been a focus, since there is a prospect that such studies could facilitate the breeding and genetic engineering of crop species, thus improve the yield and quality of our food resources. The grass family Poaceae is arguably one of the most important plant families, feeding billions of people by providing starch from wheat, maize, rice, millet and sorghum. Many important industrial materials, such as fiber and sugar, are also from Poaceae species (bamboo and sugarcane). Massive efforts have been made for the breeding of above mentioned Poaceae crop species, and significant improvements have been achieved in the yield and stress-resistance of crops. However, the phylogeny of Poaceae is not easy to resolve, because this family has a complex evolutionary history and a large size with over 11,000 species. Currently, the family is divided into twelve subfamilies, namely Anomochlooideae, Aristidoideae, Arundinoideae, Bambusoideae (bamboos), Chloridoideae, Danthonioideae, Micrairoideae, Oryzoideae (rice), Panicoideae (maize and sorghum), Pharoideae, Pooideae (wheat and barley) and Puelioideae. Although the monophyly (all and only members of a certain groups trace back to a common ancestor) of most subfamilies are well-supported, uncertainties remain among smaller groups. For instance, the relationships among tribes and sub-tribes within Panicoideae, Pooideae and Chloridoideae are still not fully resolved. Also, the current phylogeny lacks support from

nuclear genes for some lineages, since previous studies were largely based on chloroplast genes which could only partially reflect the evolutionary history.

To obtain a comprehensive phylogeny of Poaceae, my sampling covers all twelve subfamilies, most tribes (45/53) and 231 genera. *De novo* assembly of 342 transcriptomes and seven low-depth genomes were performed on RNA/DNA sequencing data from both fresh and herbarium samples. For molecular marker, I chose nuclear genes which are currently widely utilized to resolve species phylogeny and can be easily obtained from sequencing data. To minimize confounding signals from paralogs, I selected low-copy putative orthologous nuclear genes that are less prone to this defect. 1,234 candidate genes are selected from ten representative species and used to retrieve homologs from all the transcriptomic/genomic data sets. Coalescent and super-matrix analyses were performed on subsets with hundreds of orthologous genes to estimate the Poaceae phylogeny. Results strongly support the monophyly of eleven subfamilies; however, the subfamily Puelioideae was split into two non-sister clades, one for each of the two previously defined tribes, supporting a hypothesis that places each tribe in a separate subfamily. As an extension, phylogeny of the subfamily Panicoideae was further refined by expanded sampling with additional genome skimming and transcriptome datasets, covering eleven out of fourteen Panicoideae tribes. Results supported monophyly of most tribes and discovered novel relationships.

Besides the relationship among species, another aspect of phylogeny is the divergence time. Using 180 nuclear genes and 13 fossil records as calibrations, molecular clock analyses estimated the crown age of Poaceae to be ~101 million years (my; this node was fixed). Following the successive divergences of the basal subfamilies spanning a period of ~20 my, the crown age of (PACMAD + BOP) is estimated to be ~81 my in the Upper Cretaceous. Thus, the PACMAD and BOP clades probably diverged before the Cretaceous-Paleogene (K-Pg) boundary.

Plants can be classified as $C_3$, $C_4$ or CAM (Crassulacean acid metabolism) based on the type of photosynthetic pathway, and in Poaceae there are grasses of both $C_3$ and $C_4$ types, making

it a good system to study the evolution of photosynthesis. $C_4$ species are better adapted to environment with less precipitation and higher temperature and are more efficient in assimilating atmospheric $CO_2$ in such conditions. Therefore, $C_4$ crops such as maize and sorghum are generally more resistant to stress and have higher yield, compared with $C_3$ crops such as rice and wheat. To improve the yield of rice, scientists have been studying the genetic basis of $C_4$ photosynthesis with the expectation to convert $C_3$ rice into $C_4$ (see the $C_4$ rice project https://c4rice.com/). To better understand the evolution of photosynthesis, a well-resolved Poaceae phylogeny can serve as a basis. In my project, based on the information of photosynthesis type ($C_3$/$C_4$) of grass species, the ancestral states of lineages are reconstructed by parsimony method. Results support a hypothesis of multiple (at least five based on my results) independent origins of $C_4$ photosynthesis. This is further supported by phylogenetic analysis of the *ppc* gene family suggesting the recruitment of members from three paralogous subclades (*ppc-aL1a*, *ppc-aL1b*, and *ppc-B2*) as functional $C_4$ *ppc* genes.

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# ACKNOWLEDGEMENTS

I sincerely thank my wife *Ling Tang* for her love and support during my time spent here in State College. Thank my parents for their advice on my career development and their financial support.

I earnestly thank Dr. Ma for mentoring me through the years. He is always enthusiastic, optimistic and realistic about research.

To all current and previous lab members: it was great to work and enjoy the picnics with you! I learnt a lot from you, not only about science, but also about how to balance work and life.

Lastly, I would like to thank Weis, Walmart, Aldi, Sam's; Big Bowl Noodle House; Penn State Campus recreation, East Coast Fitness; and all the stores and restaurants that I have been to. Thanks for making a colorful and wonderful time for me in State College.

My dissertation project was supported by: funds from Eberly College of Science and the Huck Institutes of the Life Sciences at the Pennsylvania State University; grants from the National Natural Science Foundation of China (31770242 and 31970224); funds from the Ministry of Education Key Laboratory of Biodiversity Science and Ecological Engineering and State Key Laboratory of Genetic Engineering at Fudan University.

# Chapter 1

# Introduction

In this chapter, general information and current knowledge about phylogeny of grass family Poaceae is introduced. As a technique/methodology supporting this project, the application of large-scale transcriptomic/genomic data in phylogenetics is introduced, exemplified by studies on angiosperms. Lastly, basic knowledge of photosynthetic pathway type is covered, with emphasis on the evolution of $C_4$ in plants.

## 1.1 General knowledge and phylogeny of grass family Poaceae

### 1.1.1 General knowledge of Poaceae

The grass family Poaceae (Monocots-Commelinids-Poales-Poaceae, see Figure 1-1; also called Gramineae) is widely distributed and is the fifth largest plant family, consisting of twelve subfamilies and over 11,000 species (Kellogg, 2015; Christenhusz and Byng, 2016; Soreng et al., 2017). Species from this family are commonly referred to as “grasses”, although sometimes other monocots may share this name. Some Poaceae species, such as rice, wheat, maize, millet, sorghum, and barley are domesticated by humans as major sources of staple food. Others for example bluegrass, sugarcane, bamboos, and reeds, are important fodder and forage for farm animals and industrial materials. Grasses are essential components of diverse ecosystems, including forest, grassland, wetland, and savanna. Furthermore, one of the advantageous characteristics of many grass species, including maize, sorghum, millet and sugarcane, is carbon fixation via the $C_4$ photosynthetic pathway, which involves a four-carbon intermediate, unlike the typical $C_3$

photosynthesis that uses a three-carbon intermediate (Sage, 2004; Christin et al., 2007a; Muhaidat et al., 2007; Schlüter and Weber, 2020). $C_4$ photosynthesis increases local concentration of $CO_2$ near the carbon-fixing RuBisCo enzyme, thereby improving the efficiency of photosynthesis and increasing the adaptability of $C_4$ plants, especially in hot and dry environments (Christin et al., 2007a; Edwards and Still, 2008).

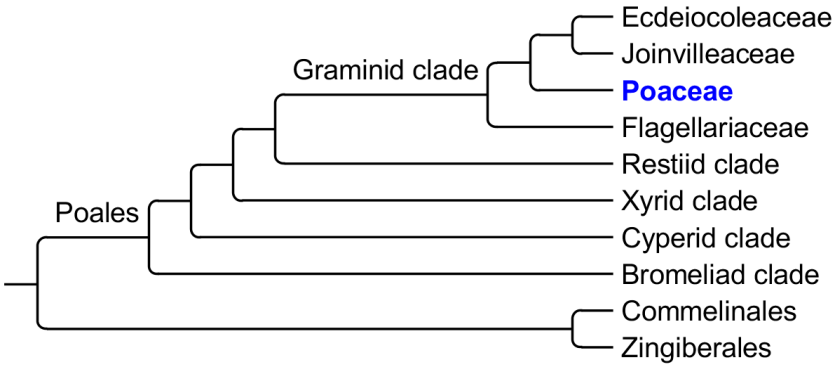


Figure 1-1: Position of Poaceae (marked in blue) in Poales. Summarized from Marchant and Briggs (2007), Bouchenak-Khelladi et al. (2014), Briggs et al. (2014) and Barrett et al. (2016).

Grasses are characterized by their usually long, narrow leaves with parallel veins. Grass stems are usually round, with internodes where leaves are attached, distinct from sedges (Cyperaceae) with triangular stems and no nodes. The inflorescence is usually compound, in which the terminal unit is in fact an unbranched cluster of flowers and each of these clusters is called a spikelet. Grass flowers are often tiny, inconspicuous and lack nectar, except for some primitive lineages (e.g., Anomochlooideae). Therefore, most grasses are pollinated by wind. The fruit of Poaceae is called a caryopsis, it is usually a dry fruit with considerable amount of starch (for example corn, rice and wheat).

The Poaceae family likely originated from moist, shady environments, as members of the basal subfamilies (Anomochlooideae, Pharoideae and Puelioideae) are mostly found in tropic forest floors. Evidence from macrofossils, pollen and phytolith (silica body) indicate the origin of Poaceae to be around 80 million years ago in the upper Cretaceous (GPWG II, 2012). Followed global

climate change, especially a cooling trend, grasses expended to open habitats, occupied new niches and developed adaptations (Estep et al., 2014; Schubert et al., 2019; Zhang et al., 2022). Bamboos (Bambusoideae) largely maintained their perennial and evergreen habit, expended further into temperate regions. Members of subfamily Pooideae are adapted to lower temperatures, with *Deschampsia antarctica* even reached Antarctica. Panicoideae diverged even more, with a smaller portion stayed in humid environments and a larger portion adapted to more arid habitats, contributing to tropical and subtropical grasslands.

**Table 1-1. Basic information of twelve Poaceae subfamilies**

| **subfamily** | **Number of species** | **Representative species** |
|---|---|---|
| Anomochlooideae | 4 | *Streptochaeta angustifolia* |
| Pharoideae | 12 | *Pharus latifolius* |
| Puelioideae | 11 | *Guaduella oblonga* |
| Bambusoideae | 1441 | *Phyllostachys edulis* (mōsō bamboo) |
| Oryzoideae | 112 | *Oryza sativa* (rice) |
| Pooideae | 3850 | *Triticum aestivum* (bread wheat), *Hordeum vulgare* (barley) |
| Panicoideae | 3316 | *Zea mays* (maize/corn), *Sorghum bicolor* (sorghum) |
| Aristidoideae | 365 | *Stipagrostis hirtigluma* |
| Chloridoideae | 1721 | *Eragrostis tef* |
| Micrairoideae | 188 | *Isachne pulchella* |
| Arundinoideae | 46 | *Phragmites australis* (reed) |
| Danthonioideae | 281 | *Danthonia spicata* |

*Species number follows Kellogg (2015).

Common species of Poaceae have long been recognized and domesticated by people around the world (see examples in Table 1-1). **Maize** (*Zea mays*, Panicoideae) was domesticated

by indigenous people in North America around 10,000 years ago (Benz, 2001), and is now the most produced cereal worldwide, with more than a dozen varieties. **Bread wheat** (*Triticum aestivum*, Pooideae) was cultivated in the regions of Fertile Crescent about 11,600 years ago, according to archaeological records (Feldman and Kislev, 2007). As the most widely consumed staple food that feeds about half of human population, **rice** (multiple *Oryza* species, Oryzoideae) was independently domesticated in China (*Oryza sativa*, 13,500~8,200 years ago; Zhang et al., 2012) and Africa (*Oryza glaberrima*, 3,500~3,000 years ago; Choi et al., 2019). Other examples of common Poaceae species include sorghum (*Sorghum bicolor*, Panicoideae), millet (multiple species in Chloridoideae and Panicoideae), and reed (*Phragmites australis*, Arundinoideae).

### 1.1.2 Current phylogeny of Poaceae

The current grass classification is built on extensive analyses of the phenetic taxonomy (mainly based on morphology) by Clayton and Renvoize (1986), Tzvelev (1989), Watson and Dallwitz (1992), Clayton et al. (2006) and their subsequent works. More recently, molecular phylogenetic analyses have facilitated revision of the Poaceae classification, leading to the current division of twelve subfamilies and molecular phylogenies of large subfamilies such as Chloridoideae (~1,700 species), Pooideae (~3,900 species) and Panicoideae (~3,300 species) (GPWG, 2001; Simon, 2007; GPWG II, 2012; Kellogg, 2015; Soreng et al., 2015; Soreng et al., 2017).

Among the twelve subfamilies, Anomochlooideae, Pharoideae and Puelioideae are small subfamilies with four, twelve, and eleven species, respectively (Clark and Judziewicz, 1996; Clark et al., 2000; Kellogg, 2015; Soreng et al., 2017), and form a grade of successive lineages sister to the remainder of the family (Figure 1-2). The other nine subfamilies form two large sister clades: the BOP clade with Bambusoideae, Oryzoideae and Pooideae and the PACMAD clade with

Panicoideae, Aristidoideae, Chloridoideae, Micrairoideae, Arundinoideae and Danthonioideae. All $C_4$ grasses are found in subfamilies belonging to the PACMAD clade, whereas members of the BOP clade, with important crops, such as rice and wheat, as well as bamboos, are all $C_3$ plants.

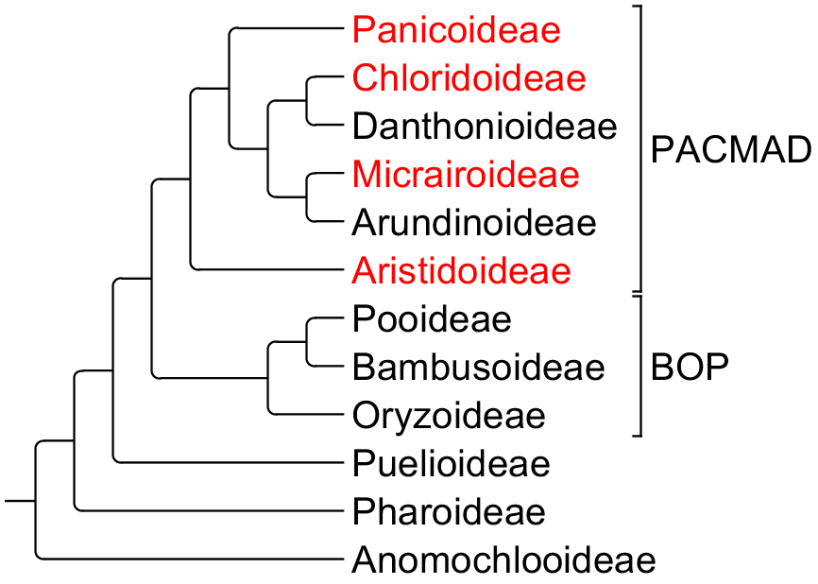


Figure 1-2. Current phylogeny of Poaceae. Relationship among subfamilies is shown, based on GPWG II (2012). Subfamilies in red contain $C_4$ species, and others are entirely $C_3$ (according to information summarized by Soreng et al. 2017).

***Phylogeny of basal subfamilies***

With dozens of published studies on Poaceae phylogeny, there are still uncertainties among Poaceae lineages, including those among subfamilies, tribes and subtribes. The early-divergent subfamilies (Anomochlooideae, Pharoideae and Puelioideae) are small and their monophyly and relationships are supported by morphological characters and phylogenies using mainly chloroplast genes (Clark and Judziewicz, 1996; Clark et al., 2000; GPWG II, 2012; Saarela et al., 2018). They are usually not sampled extensively in terms of the number of taxa and genes in phylogenetic studies, probably due to the fact that samples are hard to acquire. In the paper that first described

the subfamily Puelioideae (Clark et al., 2000), all genera of Puelioideae (*Puelia* and *Guaduella*) and Anomochlooideae (*Anomochloa* and *Streptochaeta*) were sampled, but only one out of three genera in Pharoideae were included (*Pharus*), and the result was only based on three markers (*ndhF*, *rbcL* and *PHYB*). In the study conducted by Bouchenak-Khelladi et al. (2008), not all three subfamilies were sampled for each of the three plastid markers (*rbcL, matK and trnL-F*), and the subfamily Anomochlooideae was not always monophyletic in their results. The study by GPWG II (2012) also confirmed the successive relationships of the three basal subfamilies, but in Puelioideae (with two genera, *Puelia* and *Guaduella*) they only sampled the genus *Puelia* and the analyses were only based on three plastid markers (*rbcL*, *ndhF* and *trnK/matK*). Jones et al. (2014) did a larger sampling on both Anomochlooideae (two species in *Streptochaeta* and one species in *Anomochloa*) and Pharoideae (three species in *Pharus* and one species in *Leptaspis*), although only included *Puelia* from Puelioideae, for the phylogenetic reconstruction based on four plastid loci. Their result confirmed the monophyly of Anomochlooideae and Pharoideae. In a whole plastome phylogeny by Saarela et al. (2018), subfamily Pharoideae and Anomochlooideae were relatively well-sampled, but Puelioideae was still only represented by the genus *Puelia*.

Indeed, the monophyly of the three basal subfamilies and the successive sister-relationship of them are supported by the above cited studies. However, these studies were restricted to mainly plastid marker. Considering the limitations of plastid genes, nuclear genes are needed to verify the current phylogeny.

***Phylogeny of BOP clade***

As for the BOP clade (or BEP clade, depending on the name for the rice subfamily, either Oryzoideae or Ehrhartoideae), all possible topologies among the three subfamilies have been proposed. In fact, the monophyly of this whole group was poorly supported in early phylogenetic

analyses, probably due to limited sampling. However, results from recent studies strongly support the monophyly of this clade and the (Oryzoideae, (Bambusoideae, Pooideae)) topology. For instance, Bouchenak-Khelladi et al. (2008) resolved the sister relationship of Oryzoideae to (Bambusoideae + Pooideae) by three plastid DNA regions (*rbcL*, *matK* and *trnL-F*) with a sampling covering 42% of grass genera. Edwards and Smith (2010) revealed the (O, (B, P)) topology with eight genetic markers from 1,230 grass taxa. Similarly, this topology is confirmed by GPWG II (2012) and Saarela et al. (2018) with extensive sampling of Poaceae. The above-mentioned studies were largely based on plastid genes. As for nuclear genes, Zhao et al. (2013) used 121 nuclear genes from 17 species and verified the (O, (B, P) topology. Therefore, the relationships between BOP subfamilies are relatively well-resolved. Nevertheless, the phylogeny within each of Bambusoideae, Oryzoideae and Pooideae remains to be studied.

Bamboos usually have longer generation time, and mainly reproduce asexually, which makes them evolve slower than other grasses. Using three nuclear loci (*pvcel1*, *gpa1* and *pabp1*) from 38 bamboo species, Triplett et al. (2014) revealed the herbaceous bamboo tribe Olyreae to be sister to (Arundinarieae + Bambuseae) and identified six ancestral genome donors for contemporary bamboo lineages which are hybrids. They proposed that the complex history of reticulated evolution could lead to the difficulty in bamboo phylogeny.

As the largest grass subfamily with ca. 4,000 species in 15 tribes, Pooideae has been studied but mainly based on chloroplast genes, and the relationship between tribes remains to be resolved. The rice subfamily Oryzoideae is relatively small and has been studied extensively, especially the genus *Oryza* (Zhu and Ge, 2005; Zou et al., 2008; Kumagai et al., 2010) because of its importance in agriculture. Yet the relationship among tribes and between *Oryza* and *Leersia* need to be confirmed by more extensive sampling of nuclear genes.

***Phylogeny of PACMAD clade***

The relationships among subfamilies in the PACMAD clade are incongruent among different studies with various sample size and different molecular markers and methods employed. In some early molecular phylogenetic studies, not all the six subfamilies in the PACMAD clade were yet recognized. For example, in the study conducted by Davis and Soreng (1993), only Panicoideae, Arundinoideae, Chloridoideae and Centothecoideae (now merged into Panicoideae) were recognized, and Panicoideae was estimated to be sister to the other three, by analyses of chloroplast DNA restriction site variations among 31 accessions of grasses and *Joinvillea* (Poales-Joinvilleaceae) as an outgroup. In more recent studies that recognized all the six subfamilies, there still remain problems such as which one is the basal-most (i.e., sister to the rest) in the PACMAD and the uncertain position of Micrairoideae. Bouchenak-Khelladi et al. (2008) sampled all twelve subfamilies and inferred a sister relationship between Panicoideae and the other five subfamilies in PACMAD, based on three plastid markers (*rbcL*, *matK* and *trnL-F*) using Most-Parsimony and Bayesian methods, although the relationships between the latter five subfamilies were not well-resolved. In a 1,230-taxon phylogeny that covered all BOP and PACMAD subfamilies except Arundinoideae by Edwards et al. (2010), a topology was estimated as (Panicoideae + Micrairioideae) being sister to (Aristidoideae + (Danthonioideae+ Chloridoideae)). Although six plastid and two nuclear regions were utilized for phylogenetic analyses, this study was impacted by incomplete sampling (missing basal subfamilies and Arundinoideae).

In the study by GPWG II (2012) based on three chloroplast markers (*rbcL*, *ndhF* and *trnK/matK*), Aristidoideae was estimated to be sister to the rest of the PACMAD clade, and Micrairoideae is sister to Arundinoideae. Even though, they also showed that SH tests (Shimodaira and Hasegawa, 1999) could not reject other alternative topologies for the PACMAD clade. Cotton et al. (2015) used plastome sequences from 18 species to estimate the PACMAD topology. Most-

Likelihood and Bayesian Inference analyses estimated Panicoideae to be sister to the rest of the PACMAD clade, while the result from MP analysis supports Aristidoideae to be the basal-most one in the PACMAD clade. In the whole-plastome ML analyses under different data partitions by Saarela et al. (2018), three topologies for the PACMAD clade were identified: (1) Panicoideae is sister to the rest of the PACMAD (2) Aristidoideae is sister to the rest of the PACMAD or (3) Aristidoideae and Panicoideae form a clade that is together sister to the rest of the PACMAD. There are more studies to mention, but inconsistence exists among their results. To sum up, Panicoideae, Aristidoideae and (Panicoideae + Aristidoideae) have all been proposed to be the basal lineage in PACMAD clade, and the relationships between the other four subfamilies are also unstable. This indicates that the PACMAD phylogeny is sensitive to sampling size and methods, and that there is incongruence between the evolutionary histories of different genes. To explain the elusive relationships in the PACMAD clade, rapid radiation has been proposed (Cotton et al., 2015). In such a scenario, information from plastid genes could be limited and insufficient, because the plastome is only maternally inherited and generally evolves at a slower rate. Teisher et al. (2017) compared their results from different data partitions and methods based on 131 full plastomes across the Poaceae family and indicated that the placement of basal lineage in PACMAD is unlikely to be resolved by plastome sequences.

Moreover, the phylogeny on the tribal and sub-tribal level of the large subfamilies Chloridoideae and Panicoideae is still incomplete. The subfamily Chloridoideae, for example, contains five tribes according to some studies, including the tribe Centropodieae with *Centropodia* and *Ellisochloa* (Peterson et al., 2011; GPWG II, 2012; Soreng et al., 2017); however, Fisher et al. (2016) placed Centropodieae closer to other members of the PACMAD clade rather than Chloridoideae, and thus not regarded it as a tribe in Chloridoideae. For Cynodonteae, the largest tribe in Chloridoideae, the relationships among 21 subtribes remain to be resolved. The largest PACMAD subfamily, Panicoideae, contains fourteen tribes and is diverse in morphology and other

important traits, but the relationships among Panicoideae tribes are not consistent among previous studies. For example, the relationships among early-divergent tribes in Panicoideae, including the $C_3$ tribes Centotheceae, Chasmanthieae, and Thysanolaeneae, as well as the $C_4$ tribe Tristachyideae (GPWG II, 2012; Saarela et al., 2018), deserve further investigation.

A well-supported Poaceae phylogeny can facilitate evolutionary and comparative studies, such as the evolution of inflorescence structure (Vegetti and Anton, 1995; Perreta et al., 2009) and the origin of $C_4$ photosynthetic pathways (Vicentini et al., 2008; Christin and Besnard, 2009; GPWG II, 2012). All known $C_4$ grasses are members of four subfamilies in the PACMAD clade, namely Aristidoideae, Chloridoideae, Micrairoideae, and Panicoideae. Panicoideae contains the largest number of $C_4$ grass species as well as some $C_3$ lineages (Sinha and Kellogg, 1996; Kellogg, 2015; Soreng et al., 2017). Aristidoideae, although a smaller subfamily, is also a mixture of $C_3$ and $C_4$ grasses. Previous studies have proposed multiple origins of $C_4$ photosynthesis in grasses (Sinha and Kellogg, 1996; Christin et al., 2007; Edwards and Still, 2008; Christin et al., 2012; GPWG II, 2012). However, the uncertain relationships among PACMAD subfamilies and among some lineages within Chloridoideae and Panicoideae need to be resolved to further understand the evolution of photosynthetic pathway in Poaceae.

### 1.2 Application of large-scale nuclear data in plant phylogenetics

One of the fundamental challenges in phylogenetics is to select appropriate molecular markers (genes). In theory, genes that are orthologous and not under strong selection are good candidates. There are a couple of reasons. First, phylogeny of genes does not always align with that of species, but phylogeny based on orthologous genes are more likely to reflect the species phylogeny. Second, genes that evolve largely free of selection are supposed to deviate less from the assumption of neutral mutations, and thus again fits better to resolve species phylogeny.

Such genes are supposed to be found in most if not all species of interest, and sequence information can be obtained via PCR, target enrichment, or genome/transcriptome sequencing. When the target group is large, more genes are usually required for a higher resolution. In such cases, designing primers for hundreds of genes becomes difficult. Target enrichment method, on the other hand, aims to acquire specific genomic regions that are widely conserved across taxonomic groups (for example, among angiosperm) by using baits (primers that are designed to have the right level of specificity across species). This reduces the labor to design primers for each study but also result in a higher missing rate in the output dataset. With the advancement of sequencing technology, the cost of genome/transcriptome sequencing is going down to ~1000 $ per library (shallow), making it a promising strategy for functional genetics and phylogenetic studies.

Housekeeping genes (constitutive genes, or maintenance genes) are those that required for the maintenance of basic cellular functions. They are usually expressed across different organs, cells and tissues (Butte et al., 2001; Zhu et al., 2008). Although this term was originally used for intra-specific studies in model organisms, it can be generalized for phylogenetics to refer to the group of homologous genes that can be found in the relatives of model organisms (e.g., homologs of rice housekeeping genes in grasses). For example, *rbcL* gene, which encodes large subunit of ribulose bisphosphate carboxylase, is widely expressed in photosynthetic tissues at a high level in most plants. These house-keeping genes, especially nuclear genes, are good candidates for phylogenetic studies. With an expanded pool of genes (compared with using only plastome), one can choose those that meet the criteria: low copy, (putatively) orthologous, and not under strong selection. Indeed, recent efforts using low copy nuclear genes have proven successful in resolving previously difficult relationships in large families or among more divergent lineages (Wickett et al., 2014; Zeng et al., 2014; Huang et al., 2016; Xiang et al., 2017; Yang et al., 2018; Mandel et al., 2019; Leenbens-Mack et al., 2019; Zhao et al., 2021; Zhang et al., 2022; Timilsena et al., 2022).

For the purpose of phylogenetic analyses, one does not need to be stringent about material used for transcriptome sequencing. Fresh plant tissues from vegetative organs such as leaves, stems and roots, and reproductive organs such as flowers and fruits can all be used (Huang et al., 2016a; Huang et al., 2016b; Xiang et al., 2017; Zhao et al., 2021). It is suggested to include as many types of tissues as possible, to increase the probability of obtaining genes expressed in different tissues. Fresh tissues are usually frozen or put in silica gels to dehydrate shortly after collection to reduce RNA degradation, and RNA extraction is then performed using kits. The RNA sample is then sent for sequencing on commercial platforms, and usually a certain amount of data (e.g., measured in gigabyte/gigabase) is produced on request. Therefore, it is also imperative to make sure your samples are not contaminated or degraded; otherwise, you would be paying for low quality or unless data (in terms of phylogenetic analysis).

The raw sequencing data would first need to go through quality check using some statistical package along with visualization tools (for example, FastQC; Andrews et al., 2015). A dozen measurements can be reported, but we want to pay special attention to quality scores across bases and adapter content. For both single and paired-end sequencing, adapters are used in the process, and are usually present at the ends of sequencing reads. If adapter remains are detected, trimming of data needs to be performed. There are software packages (for example, Trimmomatic; Bolger et al., 2014) that can remove low-quality bases and adapters and output the trimmed sequences. For most transcriptomes, *de novo* assembly (e.g., using Trinity; Grabherr et al., 2011) is then performed on the trimmed datasets. This step is the most time consuming and requires higher level computing resources, so using computer clusters on a paid server or public bioinformatic platforms (e.g., Galaxy: usegalaxy.org) is suggested.

Depending on the sequencing library, the assembled transcripts may contain mRNA, non-coding RNA and other types of RNA sequences. To obtain coding sequences of protein-coding

genes, one can use prediction methods (such as TransDecoder). The next step would be to select genes that are suitable for phylogenetic analyses: low-copy (ideally single-copy) and orthologs found among target species. Usually there are genes previously studied that meet these criteria, but one can also use clustering methods (such as OrthoMCL; Li et al., 2003) to obtain more (usually hundreds or thousands of) orthologous groups from species. An orthologous group is a group of genes from multiple species that are supposed to be orthologous to each other. When there are hundreds of species, running clustering on all of them would be impractical, so in that case representative species that cover major lineages of interest can be selected, and orthologous groups be predicted based on these species. Once the target orthologous groups are determined, orthologs from additional species can be obtained by BLAST-based methods (such as HaMStR; Ebersberger et al., 2009). It is possible that multiple sequences are reported for each gene, and these could be from different isoforms of the same gene, or from different copies (generated by duplication) of the same gene. At this point, objective criteria can be used to purge sequences clustered into the same orthologous group. For example, one can require only keeping the sequence of the highest similarity. However, it is possible that non-orthologs remain after this step, and this needs inspection in later steps.

Once orthologs are gathered from target species (depending on the quality of data and size of the taxonomic group of interest, one can use hundreds to thousands of genes), gene trees can be reconstructed. The sequences need to be aligned (there are several software available, for example MAFFT; Katoh et al., 2009), and then a gene tree reconstructed for each gene. Coalescent species trees can then be estimated using gene trees as input. Prior knowledge, for instance, the monophyly of some well-studies lineages (subfamily, tribe, or even genus) is required to check coalescent results (Zhao et al., 2021). This acts as a positive control to verify the validity of data. There will be cases where the new results differ from previous studies, and that would require further investigation to tell whether it reflects true biological significance or is caused by confounding

factors such as inclusion of paralogs and/or higher rate of missing data. As the evolutionary history of organisms is largely based on inference and our understanding is ever changing more data are gathered, one can only reach a point where he/she could claim that most errors and noise are eliminated, and the results reflect a scenario likely to be reliable to the best of knowledge.

In the field of phylogenetics, pioneering studies started with nuclear genes from genomic data. In the study by Liu et al. (2014), phylogenetic analyses from two low-copy nuclear genes (p*epc4* and *GBSSI*) support the recognition of three distinct subgenera in *Sorghum* (Panicoideae). Zhang et al. (2012b) identified 1,083 highly conserved low-copy nuclear genes across seven angiosperm species, then used five of them from 94 plant species and reconstructed a well-resolved phylogeny. Zeng et al. (2014) obtained transcriptomes from 26 angiosperm species representing five groups (eudicots, monocots, magnoliids, Chloranthaceae and Ceratophyllaceae), and resolved deep relationships using 59 low-copy nuclear genes.

In the years followed, transcriptomic sequencing is being increasingly applied to phylogenetic studies of plants. One obvious advantage of transcriptomic data is that coding regions of genes can be readily obtained from assembled transcripts, sparing the trouble of predicting coding sequences (CDS) from genomic data. Coding sequences are directly related to the function of proteins encoded, so for housekeeping genes with very similar function among species, CDS evolve slower compared with intergenic regions and introns. To investigate the early diversification of land plants, Wickett et al. (2014) generated 92 transcriptomic datasets and included 11 genomes. They found robust support for a sister-group relationship between land plants and one group of streptophyte green algae, the Zygnematophyceae. Huang et al. (2016) obtained 113 low-copy orthologous nuclear genes from 55 Brassicaceae datasets including 32 transcriptomes. The results improved Brassicaceae phylogeny and supported convergent evolution of several morphological traits. Xiang et al. (2017) generated 125 new transcriptomic and genomic datasets and used more

than 800 nuclear genes to resolve Rosaceae phylogeny. Ancestral state reconstruction based on the phylogeny supports independent origins of fleshy fruits from dry fruit ancestors. Zeng et al. (2017) incorporated 31 transcriptomic datasets with up to 504 low-copy nuclear genes, improved deep eudicot phylogeny and estimated diversification rates. In a large-scale project by Leebens-Mack et al. (2019), transcriptomes from over 1,100 green plants (Viridiplantae) were sequenced, and a robust phylogeny were obtained based on 410 single-copy nuclear genes. Their results revealed discordance between plastid and nuclear genes and provided a framework to investigate evolutionary questions including whole-genome duplication, incomplete sorting of ancestral variation, and speciation/extinction. Zhao et al. (2021) resolved Fabaceae phylogeny using over 1,500 nuclear genes from ~400 legume species. Moreover, the authors revealed dozens of polyploidization events by gene family analyses and proposed one/two switch(es) to rhizobial nodulation (for nitrogen fixation) followed by multiple losses in Fabaceae. Tilmilsina et al. (2022) obtained ~2,000 low-copy nuclear genes from genomic and transcriptomic datasets and resolved the relationship of all twelve monocot orders, covering 72 out of 77 families. Their highly supported results from both coalescent and supermatrix analyses are largely congruent with previous studies, but strong discordance between gene trees and species tree was revealed, indicating incomplete lineage sorting associated with rapid diversification.

As we could tell from the above-mentioned examples, there is an increasing number of nuclear genes identified for plant phylogenetic studies, and this method has proven to be powerful. With a larger number of genes, the resolving power is enhanced, especially for coalescent-based methods. In addition, gene family analyses can be performed using the nuclear genes obtained, contributing to the studies of morphological traits and biochemical pathways.

As for Poaceae, during the first decade of the 21st century phylogenetic studies have largely relied on plastid and mitochondrial genes or a small number of nuclear genes (relevant studies are

reviewed in Kellogg, 2015; Soreng et al., 2015; Soreng et al., 2017), with more recent studies starting to use more nuclear genes. Focusing on the subfamily Chloridoideae, Fisher et al. (2016) used 122 nuclear genes from 47 grass species to resolve relationship among five Chloridoideae tribes. Dunning et al. (2017) reconstructed phylogeny among 37 species from the BOP and PACMAD clades using 200 single-copy genes. Using more than 150 transcriptomic/genomic datasets, Zhang et al. (2022) resolved the relationship among 15 Pooideae (a Poaceae subfamily) tribes and 24 subtribes. Utilizing the genes obtained from these datasets, they proposed that gene duplications in Pooideae might have promoted adaptation to cold habitats. These studies established a decent methodology for transcriptome-based phylogenetics in Poaceae, but a more comprehensive sampling of the whole family is needed. Further analyses using a relatively large number of genes available from the nuclear genome can potentially resolve many of the remaining questions in Poaceae phylogeny and contribute to the study of other evolutionary questions.

Indeed, recent efforts using low copy nuclear genes have proven successful in resolving previously difficult relationships in large plant families and among more divergent lineages, and a promising prospect is seen for the study of relevant questions, such as whole-genome duplication, divergence time, hybridization and evolution of specific genes. Nevertheless, the number of genes currently being used only take up a small part of the transcriptomic/genomic data space. With the development of methodology and continuous exploration, more discovery is anticipated in this field.

## 1.3 Evolution of photosynthetic pathway in Poaceae

### 1.3.1 Poaceae as a good system to study $C_4$ evolution

During photosynthesis, with the energy from light, plants synthesize carbohydrate from carbon dioxide and water. For the majority of plants, the first product during carbon dioxide fixation

is 3-phosphoglyceric acid (PGA), a 3-carbon acid (Figure 1-3) which then goes through the Calvin cycle to produce sugars. This is also called the $C_3$ pathway because of the 3-carbon compound. For some plants that inhabit arid/hot environments, a modified pathway is used, where in mesophyll cells carbon dioxide is first fixed into oxaloacetate which is a 4-carbon compound, and then converted to malic acid. Malic acid is then transferred into bundle sheath cells and broken down to release $CO_2$, which is used for the Calvin cycle. This modified version of photosynthetic pathway is thus called $C_4$ pathway, because of the 4-carbon compound. In hot and dry environments, plants tend to close stomata to retain water, leading to reduced access to atmospheric $CO_2$. With lower $CO_2$ concentration in photosynthetic tissues, the rate of photorespiration (a process where RuBisCO oxygenates RuBP) tends to be higher, leading to a waste of energy. To cope with this stress, many $C_4$ plants have evolved a specialized organization of leaf tissues called Kranz anatomy (Tregunna et al., 1970; Smith and Epstein, 1971) that could increase local concentration of $CO_2$ near the carbon-fixing enzyme Rubisco, by physically separating the light-dependent reactions and the Calvin cycle into mesophyll cells and bundle sheath cells, respectively. By increasing the concentration of $CO_2$ around Rubisco, photorespiration is reduced, and the efficiency of carbon fixation is improved. Integrating $C_4$ pathway into Kranz anatomy, $C_4$ plants have an advantage over $C_3$ plants under hotter conditions. Sage et al. (2018) proposed a model to explain the evolution from $C_3$ to $C_4$ photosynthesis, driven by a demand to refix photorespired $CO_2$. While photorespiration depresses $C_3$ performance, the photorespired $CO_2$ can be exploited to build an evolutionary bridge to $C_4$ photosynthesis. In interestingly, Voznesenskaya et al. (2001) reported $C_4$ pathway without Kranz anatomy in *Borszczowia aralocaspica*, a Chenopodiaceae species that accomplishes $C_4$ photosynthesis within the chlorenchyma cell cytoplasm. This is a good example of the different mechanism plants utilized for $C_4$ pathway and is supportive for the idea of multiple independent origins of $C_4$.

An alternative $CO_2$-concentrating mechanisms utilized by plants adapted to arid environments is the Crassulacean acid metabolism, or CAM. In plants using CAM, the stomata on leaves remain closed during daytime to reduce evapotranspiration when the temperature is high, while also blocking intake of $CO_2$. During night stomata open to collect carbon dioxide and store it as malic acid in vacuoles, which is later used to release $CO_2$ back into chloroplast under sunlight to go through the Calvin cycle. With the CAM pathway, plants can save water in arid conditions while maintaining efficient photosynthesis. CAM plants are reported among dozens of families, including Bromeliaceae, Cactaceae, Crassulaceae and Orchidaceae (Winter et al., 1983; Nobel and Hartsock, 1986; Griffiths, 1989; Crayn et al., 2004).

There are around 8,100 known $C_4$ plant species (Sage, 2016), accounting for less than 3% of flowering plants (~300,000 species, Christenhusz and Byng, 2016), but contributing to ~23% of global primary biomass production (Kellogg, 2013). Examples of $C_4$ plants include species in Amaranthaceae (e.g., *Amaranthus*), Asteraceae (e.g., *Flaveria*), Cyperaceae (e.g., *Cyperus*), Euphorbiaceae (*Euphorbia;* contains the only known $C_4$ trees) and most commonly, Poaceae, with the largest number (~4,500 species, ~60% of all $C_4$ plants) of $C_4$ species.

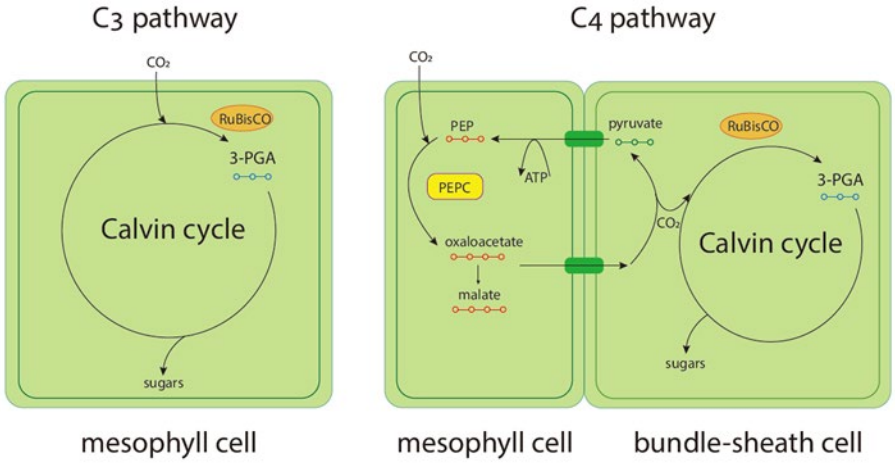


Figure 1-3 : Difference between $C_3$ and $C_4$ photosynthetic pathways. PEP: Phosphoenolpyruvate; PEPC: Phosphoenolpyruvate carboxylase; 3-PGA: 3-Phosphoglyceric acid. The enzyme PEPC is multi-functional, with the $C_4$ version distinctive from others by unique amino acid residues. Illustration was modified from

https://www.khanacademy.org/science/biology/photosynthesis-in-plants/photorespiration--c3-c4-cam-plants/a/c3-c4-and-cam-plants-agriculture.

Based on phylogenetic analysis, $C_3$ pathway is proposed to be ancestral, and $C_4$ has probably originated independently multiple times in different lineages. For example, GPWG II (2012) inferred 22~24 origins of $C_4$ in Poaceae, based on a chloroplast-gene phylogeny. McKown et al. (2005) proposed multiple $C_4$ origins even within the Asteraceae genus *Flaveria*. One piece of supporting evidence for such inference is that $C_4$ lineages are often intertwined with $C_3$ lineages, and $C_3$ lineages are usually at more ancestral positions along the phylogeny. Although multiple genes and anatomical features are involved in $C_4$ pathway (Christin et al., 2015; Moreno-Villena et al., 2018), the conversion seems to have happened frequently, making the underlying mechanism intriguing.

Grasses occupy a wide range of habitats, and $C_4$ grasses are especially dominant on tropical and subtropical grasslands (e.g., savanna). The success of $C_4$ grasses is thought to be due in part to their ability to fix carbon via $C_4$ photosynthesis, which facilitates adaptation to various niches. To date, all reported $C_4$ grasses belong to four subfamilies of the PACMAD clade, namely Aristidoideae, Micrairoideae, Panicoideae and Chloridoideae. Based on previous phylogeny, either Panicoideae, Aristidoideae or these two combined is the basal-most lineage of PACMAD clade (see Figure 2-2 for a comparison) and both of them are mixture of $C_3$ and $C_4$ species. However, the existence of undiscovered $C_4$ species in Arundinoideae or Danthonioideae cannot be excluded, and the origin of $C_4$ in PACMAD clade deserves further investigation. There are even both $C_3$ and $C_4$ ecotypes/subspecies within a single species, such as *Alloteropsis semialata* (Lundgren et al., 2016). The large number of possible $C_4$ origins and complexity of $C_4$ distribution among lineages make Poaceae an idea system to study $C_4$ evolution. A more comprehensive sampling of Poaceae species and more robust phylogeny would serve as a basis for such studies.

### 1.3.2 PEPC as a key enzyme in $C_4$ photosynthesis

Among genes involved in $C_4$ photosynthesis, *ppc* genes that encodes phosphoenolpyruvate carboxylase (PEPC; EC 4.1.1.31), which is responsible for the initial fixation of atmospheric $CO_2$, has been studied in several plant families including Poaceae, Asteraceae and Fabaceae (Bläsing et al., 2000; Christin et al., 2007b; Christin and Besnard, 2009; Wang et al., 2016). The *ppc* genes belong to a gene family encoding several similar enzymes involved in photosynthesis and stress-response processes. Found in bacteria, green algae and higher plants (but not found in animals and fungi), PEPC catalyzes the addition of bicarbonate ($HCO_3^-$) to phosphoenolpyruvate (PEP) to form oxaloacetate and inorganic phosphate (Kai et al., 2003). This reaction is used in both $C_4$ and CAM pathways.

Based on motifs, gene structure and gene family analysis, PEPCs are classified into two types: plant-type PEPC (PTPC) and bacterial-type PEPC (BTPC) (Kai et al., 2003; Sánchez and Cejudo, 2003; O'Leary et al., 2011; Wang et al., 2016). Most plants (including green algae) have at least one BTPC, and the number of *ppc* genes (counting both PTPC and BTPC) in a plant species ranges from 2 to 10 (Wang et al., 2016). With around 970 amino acid residues (PTPC), the C-terminus of PEPC is usually highly conserved among species, while the N-terminus is more variable.

The enzyme typically consists of four identical subunits (dimer of dimers), although isoform composed of three different subunits was reported in a unicellular alga (Rivoal et al., 2001). In *Zea mays*, the $C_4$ PEPC is composed of four identical subunits; it's worth noting that *Z. mays* has only one *ppc* gene that encodes the $C_4$ PEPC, while in some other grasses there are multiple $C_4$-type PEPCs. Most PEPCs are regulated by allosteric effectors. In plants, PEPCs are activated by glucose-6-phosphate (G6P) or glycine and inhibited by L-malate, aspartate or glutamate.

Regulatory phosphorylation, which could be important in plant response to light, salt stress and $CO_2$ concentration, occurs at the Serine residue (#15 in *Zea mays* PEPC) close to the N-terminus.

Three-dimensional structures of PEPC from *E. coli* and maize have been elucidated by X-ray crystallographic analysis (Matsumura et al., 2002; Kai et al., 2003), and the enzyme was revealed to be highly regulated by both phosphorylation and allostery. N968 and G970 (*Zea mays* $C_4$ PEPC numbering; NP_001154820.1) are important for allosteric regulation and catalytic activity, respectively (Dong et al., 1999). H177 is considered the most important catalytic base (Terada and Izui, 1991). R183, R184, R231 and R372 are strictly conserved in plant PEPCs and are deduced to be the binding sites for G6P which is an allosteric activator (Blasing et al., 2000).

The $C_4$-version PEPC has distinct kinetic and regulatory properties compared to non-$C_4$ ones, although they share high sequence similarity. Previous studies of the *ppc* family indicated that *ppc* genes for $C_4$ photosynthesis encode proteins with shared sequence motifs (Bläsing et al., 2000; Christin et al., 2007a; Paulus et al., 2013). In photosynthetic angiosperms, the number of *ppc* genes in a species varies from 2~10 (Wang et al., 2016). $C_4$ *ppc* genes in Poaceae originated from non-$C_4$ paralogs in two different *ppc* clades (Christin and Besnard, 2009), sometimes involving possible horizontal gene transfer (Christin et al., 2012). Specifically, a conserved Serine/Alanine residue (corresponding to residue #780 in the *Zea mays* PEPC, GRMZM2G083841) was shown to be responsible for the kinetic differences between $C_3$ and $C_4$ isoforms with respect to the substrate PEP (Bläsing et al., 2000). This residue is so far verified to be conserved in all known $C_4$ plants, including Poaceae and Asteraceae, and even in CAM plants, for example Orchids.

Current knowledge indicates there are five or six *ppc* clades in Poaceae, but earlier studies with limited sampling depicted fewer *ppc* lineages. Gehrig et al. (2001) identified three functional *ppc* isoforms (*ppc-aL, ppc-aR* and *ppc-C*$_4$) from *Zea mays*, *Sorghum vulgare*, *Saccharum sp.* and *Triticum aestivum*. With an expanded sampling, Besnard et al. (2003) revealed four *ppc* lineages

(*ppc-aL1, ppc-aL2, ppc-aR* and *ppc-C4*) and proposed that the C4 *ppc* gene is derived from *ppc-aR*, which is a highly expressed isoform in roots. They also proposed that both polyploidization and tandem duplication contribute to the expansion of *ppc* gene family. Christin et al. (2007a, b) described five *ppc* clades (*ppc-aL1, ppc-aL2, ppc-aR, ppc-B2* and *ppc-B1*) and proposed that natural selection among the C4 *ppc* genes could cause bias on the phylogeny of *ppc* gene family using all codon positions, resulting in C4 *ppc* genes clustered together as a sister to *ppc-B2*. Using only 3rd codon positions (from exons 8, 9 and 10) and introns combined, they were able to resolve the topology for the branches containing *ppc-B2* and C4 *ppc*. The authors further demonstrated that in C4 *ppc* there are 21 codons under positive selection and excluding them as they claimed could improve the *ppc* gene phylogeny to be closer to organism phylogeny. Their results show that C4 *ppc* genes are derived from within *ppc-B2* clade multiple times independently and should be an integral part of it rather than a separate clade. It's worth noting that introns and 3rd codon positions are still under selection pressure in some circumstances. For example, codon usage bias may result in higher frequency of certain codons over other synonymous alternatives, resulting in a higher frequency of GC content at 3rd codon positions; introns, on the other hand, may affect transcription dynamics and is not free of selection. Therefore, a phylogeny with introns and 3rd codons are not necessarily better than a phylogeny with all codon positions. As a follow up, Christin et al. (2009) expended sampling to include more grass species for *ppc* gene, and split ppc-*aL1* into *ppc-aL1a* and *ppc-aL1b*. Thus, for most Poaceae lineages, there are six *ppc* gene clades (*ppc-aL1a, ppc-aL1b, ppc-aL2, ppc-aR, ppc-B1* and *ppc-B2*). In their results, none of the C4 grass species have *ppc-B1*. Focused on genera *Aristida* and *Stipagrostis* in subfamily Aristidoideae, they found these two genera recruited *ppc-B2* and *ppc-aL1b* for C4 pathway, respectively, and this supports independent origins of C4 in *Aristida* and *Stipagrostis*. Also, given the fact that these two pairs of genes, *ppc-B1/B2* and *ppc-aL1a/aL1b* are each on homologous chromosome regions, they proposed that duplication might have relaxed purifying selection pressure and facilitated neofunctionalization to

be recruited for $C_4$ pathway. Interestingly, Cerros-Tlatilpa and Columbus (2009) reported a $C_3$ species, *Aristida longifolia*, that is sister to the remaining species in this genus which are all $C_4$. In a phylogenetic perspective, this is in agreement with the results from Christin et al. (2009), suggesting that $C_4$ in *Aristida* originated in the common ancestor of the remaining species (excluding *Aristida longifolia*), and that $C_4$ in *Stipagrostis* was from another independent origin.

To summarize, previous phylogenetic analysis of *ppc* gene sequences from Poaceae species, other Poales, monocots and several eudicot families helped to define six clades for grass *ppc* genes (here named as subclades): *ppc-aL1a*, *ppc-aL1b*, *ppc-aL2*, *ppc-B1*, *ppc-B2*, and *ppc-aR.* However, the origins of these subclades were not clear, whether they were present in the common ancestor of Poaceae or even Poales, or they were produced by more recent duplication events within Poaceae is a question that deserves further investigation.

# Chapter 2

# Poaceae phylogeny based on low-copy nuclear genes and evolution of $C_4$ photosynthesis

In this chapter, I report a well-resolved Poaceae phylogeny based on over one thousand low-copy nuclear genes. The relationship among subfamilies, tribes and subtribes received high support from multiple coalescent analyses. The evolution of $C_4$ photosynthesis is discussed based on the results of ancestral state reconstruction and gene family analysis of *ppc* genes.

## 2.1 Introduction and objectives

### 2.1.1 Resolve Poaceae phylogeny using low-copy nuclear genes

Poaceae is an economically and ecologically important plant family and also an ideal system to study evolutionary questions. A well-resolved phylogeny can serve as a backbone for relevant studies. In this project, I aim to resolve the family-wide phylogeny of Poaceae to the tribal/sub-tribal level, i.e., to get a phylogeny that the relationships among subfamilies and among tribes/sub-tribes within each subfamily are well-supported with no polytomy (uncertainty of relationship among multiple lineages). The genetic markers I will be using are hundreds of low-copy orthologous nuclear genes from transcriptomic and genomic data. These genes are putatively house-keeping genes, i.e., genes that are relatively conserved in terms of sequence and are sufficiently expressed in most types of plant tissues. I will include transcriptomes/genomes from both that generated by our group and from public databases. The RNA/DNA-seq datasets generated in this project will go through quality check, trimming and de-novo assembly to get contigs

(assembled sequences); contigs will be further processed to get non-redundant coding sequences (CDS). For public genomes/transcriptomes, non-redundant coding sequences are retrieved directly from NCBI database. SRA data sets will be processed the same as the transcriptome datasets generated by our own project. Putative orthologous genes will then be obtained from these CDS data sets. The genes will then get aligned automatically using software. The alignment matrices will be further trimmed to remove poorly aligned regions/sequences.

Both coalescent and super-matrix methods will be implemented to infer Poaceae phylogeny from nuclear genes. For coalescent analyses, multiple sets of genes will be selected from the pool (which totally contains 1,234 genes). Single-gene trees will be reconstructed, and the coalescent trees estimated from these different sets of single-gene trees. Statistical analyses will be used to compare the difference between single gene trees. Super-matrix tree will be reconstructed from a matrix of concatenated genes. The results from multiple analyses will be compared and summarized, and incongruence will be checked and discussed. In case of obvious conflicting results between different analyses, i.e., different topologies, explanations will be proposed. See Figure 2-1 for a workflow of the phylogenetic analyses.

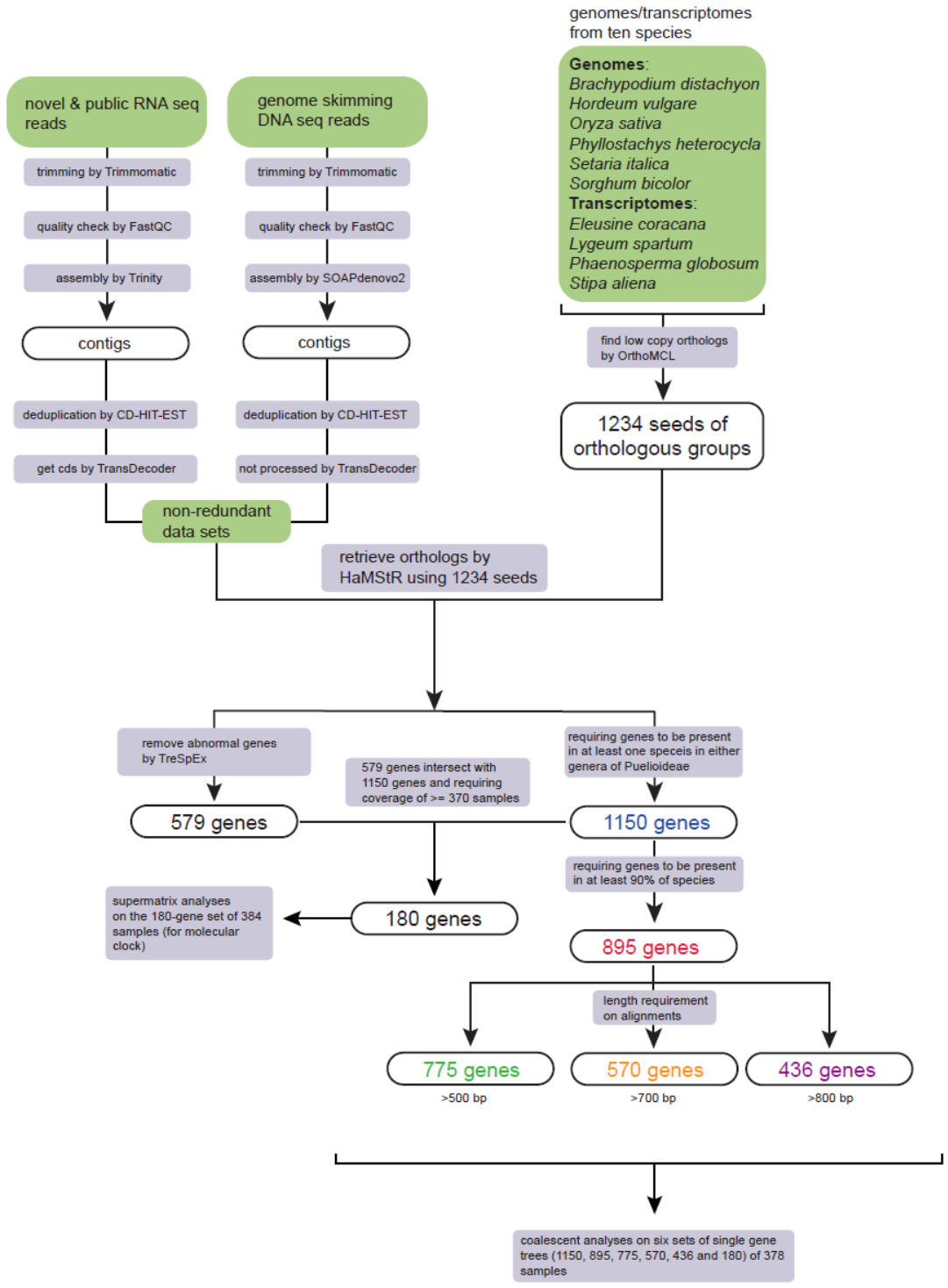


Figure 2-1: Workflow for phylogenetic analyses in this project. The gene sets 1150, 895, 775, 570 and 436 were used for coalescent analyses; gene set 180 was used for super-matrix analyses and divergence time estimation.

With the development of sequencing techniques, phylogenetic studies have progressed from using only a couple of genes to the whole plastome and then tens of nuclear genes. In earlier studies, not all Poaceae subfamilies were well supported to be monophyletic. For example, in the phylogeny by (Bouchenak-Khelladi et al., 2008), three plastid genes were unable to confirm the monophyly of Anomochlooideae, Micrairoideae or Arundinoideae, and the BOP clade relationship was estimated to be (Oryzoideae, (Bambusoideae, Pooideae)), which was incongruent with later studies. Interestingly, using two genes, *phyB* (nuclear) and *ndhF* (plastid), Vicentini et al. (2008) reported a Poaceae phylogeny that is much more improved, especially for the BOP clade and the position of Aristidoideae among the PACMAD clade, suggesting using nuclear genes could add power to phylogenetic resolution. GPWG II (2012) reported a phylogeny based on three plastid genes and is currently the most acknowledged one supported by more recent studies (e.g., by Teisher et al., 2017; Soreng et al., 2017). In this phylogeny, Anomochlooideae, Pharoideae and Puelioideae are three basal subfamilies that form a grade to the rest, or "core Poaceae". Inside the BOP clade, Oryzoideae is sister to Bambusoideae plus Pooideae. For the PACMAD clade, Aristidoideae is the basal lineage, followed by Panicoideae, and the remaining four subfamilies form two well supported groups, (Micrairoideae, Arundinoideae) and (Chloridoideae, Danthonioideae). Nevertheless, whole-plastome based phylogeny (Cotton et al., 2015; Saarela et al., 2018) tend to place Panicoideae as the first divergent one in PACMAD clade. Somewhat different from above-mentioned phylogenies is the study by Fisher et al. (2016). Using MP-EST (this method estimates species trees from a set of gene trees by maximizing a pseudo-likelihood function) with 56 nuclear genes, they reported a PACMAD topology where (Aristidoideae, Panicoideae) together is sister to the rest, and neither the (Micrairoideae, Arundinoideae) or (Chloridoideae, Danthonioideae) topology was revealed. This could possibly be explained by their methodology, as MP-EST is seldom used in other Poaceae studies. Therefore, the PACMAD topology is still open to investigation.

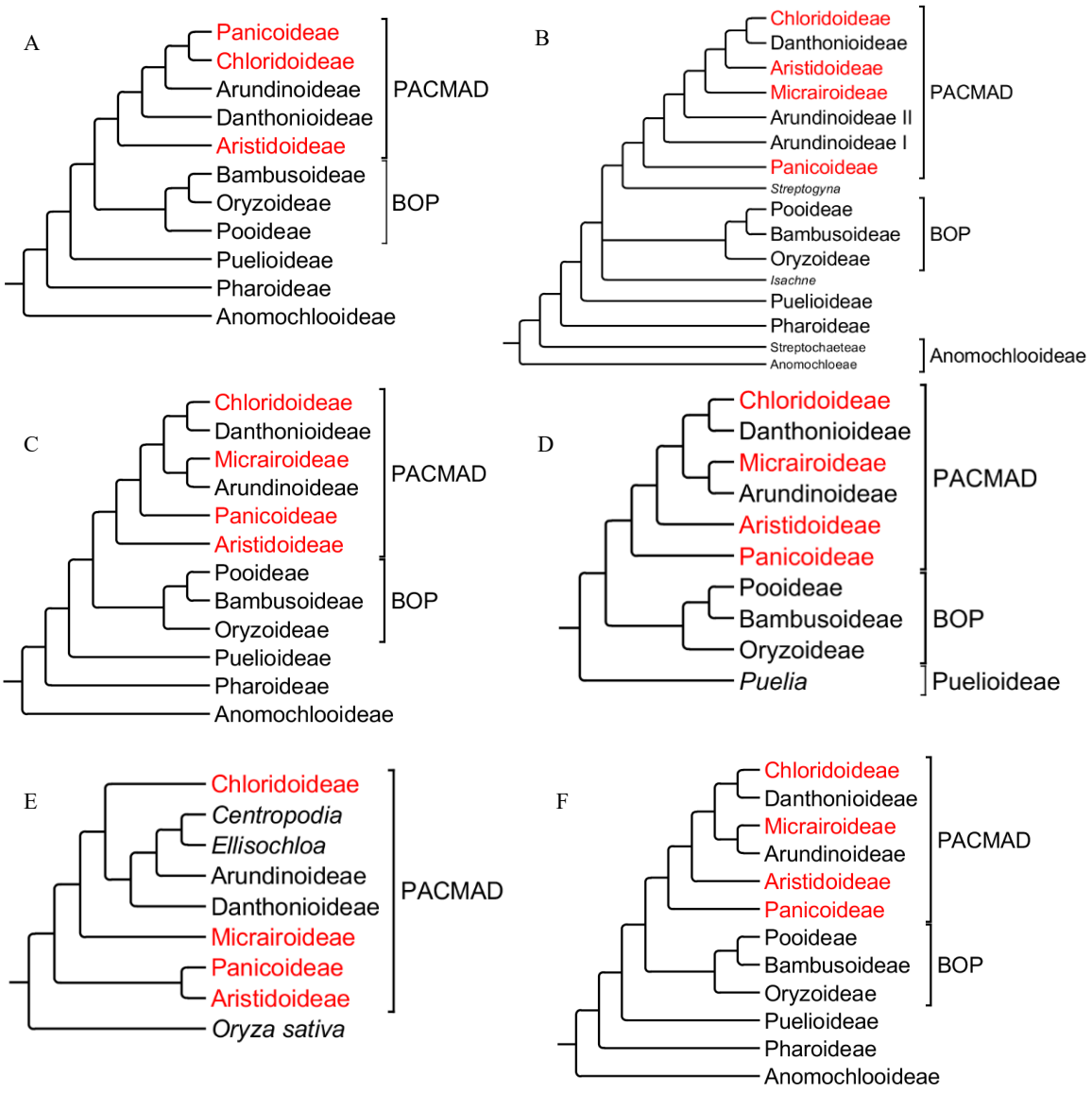


Figure 2-2: A comparison of Poaceae phylogeny estimated by previous studies. Subfamilies with $C_4$ grasses are marked by red according to Soreng et al. (2017), and single species/genera are shown by italicized fonts. (A) Vicentini et al., 2008; *ndh*F and *phy*B. (B) Bouchenak-Khelladi et al., 2008; three plastid genes. (C) Topology summarized from GPWG (2012), based on three chloroplast genes and shared by Teisher et al. (2017), based on whole plastome and also shared by Soreng et al. (2017). (D) Summarized from Cotton et al. (2015), based on whole plastome. (E) Topology based on Fisher et al. (2016), a MP-EST species tree from 56 housekeeping genes. (F) Poaceae phylogeny by Saarela et al. (2018), based on whole plastome.

Subfamily **Aristidoideae**, as well as its three genera, are supported by multiple studies to be monophyletic. *Aristida* with ~300 species is all $C_4$ except for *Aristida longifolia*, which is sister to the remainder of this genus. *Sartidia*, a smaller genus with only five species, is $C_3$ and sister to *Stipagrostis* ($C_4$). *Sartidia* and *Stipagrostis* together are sister to *Aristida*. Based on different *ppc* genes recruited for $C_4$ PEPC, Christin et al. (2009) proposed two independent $C_4$ origins in Aristidoideae. **Arundinoideae** and **Danthonioideae**, although embedded in PACMAD clade, contain only $C_3$ grasses. Problems regarding specific genera exist in Danthonioideae. *Merxmuellera rangei* and *M. papposa*, formerly in Danthonioideae, were later revealed to be closer to Chloridoideae (as tribe Centropodieae; Peterson et al. (2010). As for *Cortaderia*, evidence from morphology and molecular data suggested it might be paraphyletic (Barker et al., 2003). Subfamily Arundinoideae is now relatively small with fewer than 50 species, but a couple of genera now belong to other PACMAD lineages or even Pooideae were misplaced in this subfamily before molecular evidence was available. In recent studies, with those taxa purged Arundinoideae is now monophyletic, although species placed in other subfamilies were also reported to be in Arundinoideae, for example "*Eragrostis*" *walteri* (this genus is supposed to be in Chloridoideae; Ingram et al., 2011; GPWG II. 2012). Micrairoideae with nine genera is divided into three tribes by Soreng et al. (2017), although Kellogg (2015) argued it's unnecessary for this small subfamily. *Eriachne* and *Pheidochloa* (Tribe Eriachneae) are reported to be $C_4$, and their leaf anatomy and gene expression are different from other $C_4$ grasses, supporting an independent $C_4$ origin in this lineage.

Subfamily **Chloridoideae**, with ~1,700 species, is divided into five tribes by recent molecular phylogenies, and the relationship among them has been consistent among most studies (Figure 2-3): Centropodieae is the first diverging lineage, followed by Triraphideae, then Eragrostideae, and (Zoysieae, Cynodonteae). Centropodieae is a small tribe with only two genera, *Ellisochloa* and *Centropodia*. Although the sister relationship of these two genera are

confirmed, it has not always been treated as a lineage for this subfamily. Based on seven plastid markers, Peterson et al. (2011) resolved the relationship among over 80 Chloridoideae species, and first proposed the tribe's name Centropodieae and included these two genera into Chloridoideae. However, GPWG II (2012) did not treat Centropodieae as part of Chlridoideae, although based on their phylogeny it is clearly sister to the remaining four tribes. Fisher et al. (2016) reported a phylogeny based on nuclear genes where *Ellisochloa* and *Centropodia* are closer to Arundinoideae and Danthonioideae, directly disputing Centropodieae as a Chloridoideae tribe. Their methodology is however not comparable with most other studies, thus leaving the results questionable. Later studies, such as Saarela et al. (2018), supported the relationship of Centropodieae being sister to other Chloridoideae, but the question is still under debate, partly because *Ellisochloa* is the only $C_3$ genus if it were to be included in Chloridoideae. Nevertheless, the key to this problem is definitely to get a well-supported phylogeny and resolve the position of *Ellisochloa* and *Centropodia*. As the largest tribe, Cynodonteae has over 900 species and is supported to be monophyletic by molecular data (Columbus et al., 2007; GPWG II 2012; Peterson et al. 2010, 2011, 2014). This tribe is further divided into 21 subtribes by Soreng et al. (2017), although other authors (Peterson et al., 2014) use different subtribe names and the total number varies. Some of these subtribes, though, each only contains one genus (e.g., Triodiinae, Hilariinae in Soreng et al. 2017), and there are a few *incertae sedis* genera in Cynodonteae. Because of these facts and difference in sampling among studies, I am not making a detailed comparison for the Cynodonteae topology. It's worth noting that previous studies utilized only plastid genes or a small number of nuclear genes to resolve Chloridoideae phylogeny, and a more comprehensive sampling plus additional nuclear genes could improve the resolution.

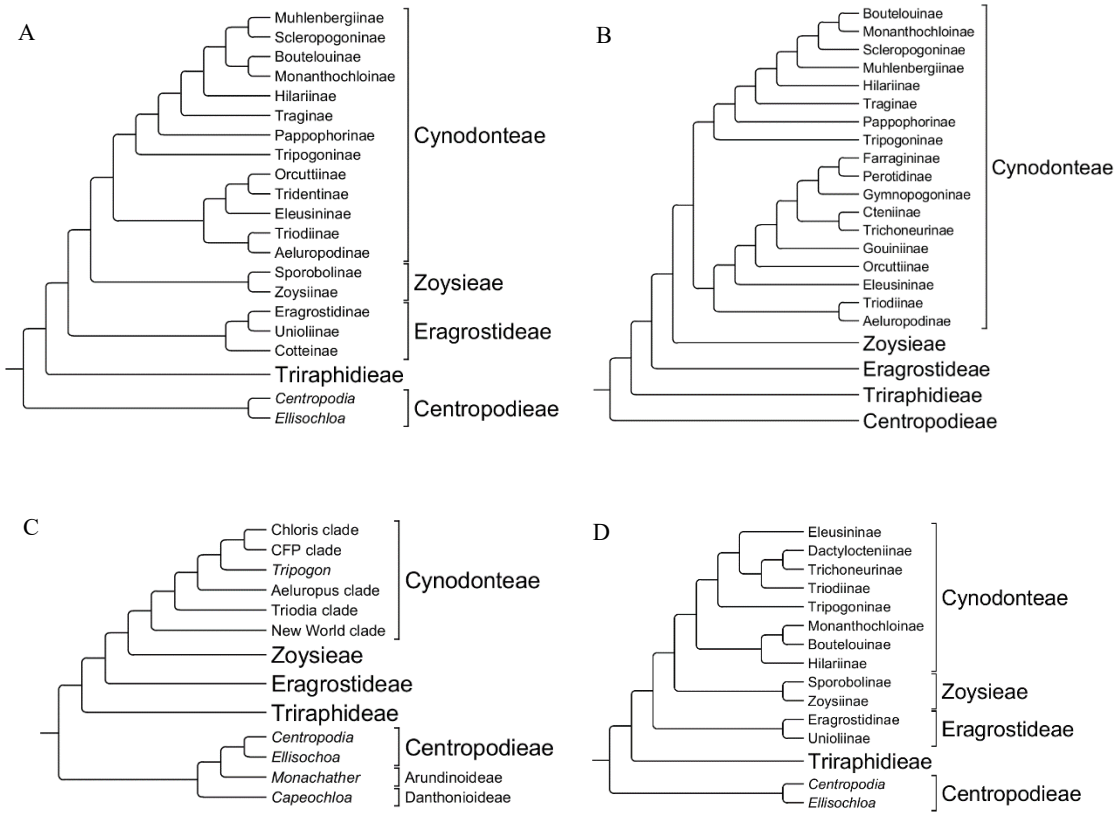


Figure 2-3: A comparison of Chloridoideae phylogeny from previous studies. (A) Summarized from Peterson et al. (2011), based on seven plastid regions. (B) Summarized from Peterson et al. (2014), using combined plastid and ITS sequences. (C) Summarized from Fisher et al. (2016), MP-EST species tree from 56 genes. (D) Summarized from Saarela et al. (2018), whole-plastome phylogeny.

First recognized by Clark et al. (1995), the monophyly of BOP clade and its three subfamilies has been verified by a dozen of phylogenetic studies with molecular data, leaving only the position of *Streptogyna* ambiguous. It is likely sister to Oryzoideae, as Saarela et al. (2018) reported in their whole-plastome phylogeny. Subfamily **Oryzoideae**, formerly called Ehrhartoideae, is the first diverging lineage of BOP clade. Soreng et al. (2017) recognized five tribes for this subfamily and propose a relationship where Streptogyneae (*Streptogyna*) is the first to diverge, followed by Ehrharteae, and (Phyllorachideae, Oryzeae), a phylogeny that is also supported by Saarela et al. (2018). As an economically important genus, the phylogeny of *Oryza* has been resolved by Zou et al. (2008) using 142 single-copy genes, although some species such as *Oryza meyeriana* was missing. Genus *Leersia* is shown to be sister to *Oryza*, but not well-sampled in phylogenetic studies. Characterized by usually woody culms (the stem) and cyclical flowering, bamboos are distinct from other grasses. Subfamily **Bambusoideae** is divided into three tribes: Arundinarieae (temperate woody bamboos), Bambuseae (tropical woody bamboos) and Olyreae (herbaceous bamboos). Based on ploidy level, Bambuseae can be further split into neotropical woody bamboos (tetraploids) and paleotropical woodybamboos (hexaploids). Plastid gene-based phylogenies placed Olyreae sister to Bambusoideae (Wysocki et al., 2015; Soreng et al., 2017; GPWG II, 2012), a scenario supporting two origins of woodiness or one origin and a loss in Olyreae. On the other hand, inclusion of nuclear genes (Triplett et al., 2014; Wysocki et al., 2016; Guo et al., 2019) revealed an alternative relationship where Olyreae is sister to (Bambuseae, Arundinarieae). Notably, woody bamboo genomes contain at least four sub-genomes, suggesting a history of reticulate evolution. Delimitation of several bamboo genus such as *Arundinaria*, *Bashania* and *Neomicrocalamus* needs further verification.

As the largest subfamily, **Pooideae** includes over 3,800 species and is economically important. Numerous studies have been conducted on genetics of wheat, barley and the recently popular model organism, *Brachypodium distachyon*. Phylogenetic studies to date have verified the

monophyly of Pooideae subfamily and many of the fifteen tribes. Relationships among the tribes have been, however, more difficult to resolve due to the fact that hybridization is common in Pooideae (Marcussen et al., 2014; Glémin et al., 2019), causing incongruence between nuclear and plastid genes and making morphological characters less powerful for classification. In most studies, tribe Brachyelytreae (three species) is reported to be the first divergent lineage, followed by Nardeae and Lygeeae, each with only one species and is sometimes combined into a single tribe Nardeae. Inconsistences exist regarding relationship among the remaining Pooideae tribes, but Brachypodieae (where *Brachypodium* belongs to) is frequently reported to be sister to a large clade containing Poeae, Triticeae, Bromeae and Littledaaleeae. Species in this clade are characterized by larger genomes compared with more basal tribes. In Triticeae which contains important crops wheat, barley and rye, maintaining nomenclatural stability has to be taken into consideration, making revisions of classification challenging. Poeae is the largest tribe in this subfamily with over 2,500 species put into more than 15 subtribes (Kellogg 2015; Soreng et al., 2017), and groups and clades were defined to further organize such a large tribe. Positions of other tribes vary among different studies, and due to the fact that these studies differed in their sampling, phylogenetic methods and molecular markers used, no conclusive description is made here. (See a comparison in Figure 2-4.)

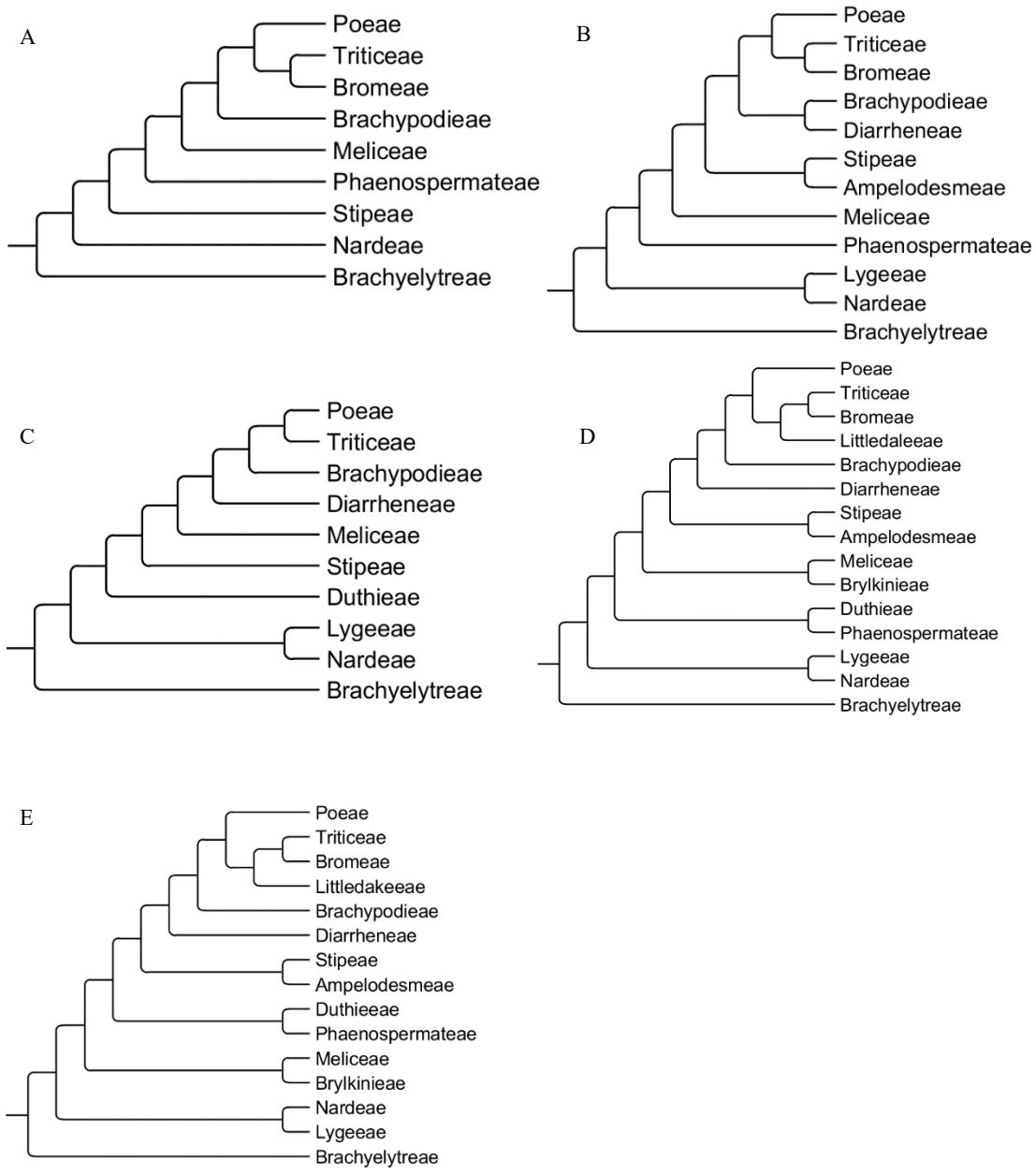


Figure 2-4: A comparison between previous phylogeny of subfamily Pooideae. Tribe names are shown. Note that these studies differ in the number of tribes sampled. (A) Summarized from GPWG (2012), based on 3 chloroplast genes. (B) Summarized from Saarela et al. (2018), based on whole plastome. (C) Summarized from Schubert et al. (2019), based on 3 chloroplast genes. (D) Summarized from Soreng et al. (2017), based on their 2015 publication with additions. (E) Summarized from Schneider et al., (2009&2011), based on nuclear and plastid DNA.

Even though previous studies have established an acceptable Poaceae phylogeny, the incongruence between them suggests that a more comprehensive sampling equipped with larger number of genes is necessary. To be more specific, earlier studies focus mostly on plastid genes, so more nuclear genes should be included, and this could potentially improve Poaceae phylogeny with the following questions to investigate:

1. PACMAD topology.

The PACMAD clade as a whole and its six subfamilies are supported as monophyletic in a number of studies, but the relationships among the subfamilies are inconclusive and sensitive to phylogenetic methods (see figure X for a comparison). Aristidoideae, Panicoideae, and these two combined have all been proposed to be sister to the reminding subfamilies of PACMAD clade. The relationship of (Chloridoideae, Danthonioideae) and (Micrairoideae, Arundinoideae) are well recognized in most studies, including A, C and D summarized in figure X. Therefore, the most outstanding question is which subfamily is the basal-most lineage of PACMAD clade.

2. Relationship of tribes and subtribes in Chloridoideae, Panicoideae, Pooideae and Bambusoideae.

For these subfamilies with a larger number of species, a better phylogeny is essential to understand evolutionary questions, such as $C_4$ evolution and origin of woodiness. The above-mentioned studies are either based on plastome genes, the whole plastome or a small number of nuclear genes. As we expand the sampling of genes to include more nuclear genes, one can expect different results for relationships among tribes and subtribes and realize that the previous studies only partly depicted the phylogeny.

### 2.1.2 Estimate the divergence time of Poaceae lineages

The divergence time of Poaceae lineages will be estimated using the branch length information (substitution rate) from the super-matrix analysis and reliable fossil records from both in-group and out-group taxa as calibrations. Bootstrapped super-matrix trees will be used to calculate the confidence intervals of node ages. The estimated node ages will be compared with major geological events such as global temperature drop/increase and change of atmospheric $CO_2$ concentration. The origin times of $C_4$ photosynthesis will also be deduced based on the ancestral state reconstruction analyses that will be described later.

### 2.1.3 Ancestral state reconstruction of photosynthetic pathway type

In Poaceae, four subfamilies are reported to have $C_4$ species, namely Aristidoideae, Micrairoideae, Panicoideae and Chloridoideae, all of which are in the PACMAD clade. Previous studies on the evolution of $C_4$ photosynthesis mainly sampled the PACMAD clade and tried to infer the number of $C_4$ origins by statistical methods. However, to reveal the genetic basis for $C_4$ photosynthesis, a more comprehensive sampling that also covers the BOP clade as well as basal subfamilies would be more helpful. For this reason, in my project I'll sample all twelve subfamilies and as many tribes as possible. I included 150 $C_4$ grasses in Poaceae (1 in Micrairoideae, 3 in Aristidoideae, 62 in Panicoideae and 84 in Chloridoideae). The designation of photosynthetic pathway type is based on information summarized by Soreng et al. (2017) on the genus level. There are a few genera that contain both $C_3$ and $C_4$ species, and a known species (*Alloteropsis semialata*) which contains both $C_3$ and $C_4$ subspecies. In such cases, the photosynthetic pathway type is confirmed by details from additional references.

To infer the ancestral state of photosynthetic pathway type, the state of 378 samples included in the coalescent analysis are coded as either 0 ($C_3$) or 1 ($C_4$), and the reconstruction analysis will be done using most parsimony method on each of the coalescent phylogenetic trees. Essentially, ancestral states ($C_3$, $C_4$ or uncertain) for all internal nodes (ancestral nodes) will be reported and mapped on the phylogenetic trees. The rationale behind maximum parsimony method is to find the least number of changes along the phylogeny, assuming that the transition rate from one state to another (or among multiple states) is constant. A famous scenario where maximum parsimony applies is inferring phylogeny from neutrally evolving nucleotide sequences, assuming the transition rate among four types of bases are equal and constant.

The assumption of constant transition rates may not hold true for all types of biological traits, so the reported result is often an underestimate. In the case of photosynthetic pathway type, the conversion from $C_3$ to $C_4$ (or reversal from $C_4$ back to $C_3$) involves multiple steps of anatomical and biochemical alterations and is inappropriate to be considered as a one-step change, not to mention to assume a constant rate of change. Also, considering the fact that $C_4$ species poses advantage over $C_3$ ones in some habitats (and vice versa), the transition from $C_3$ to $C_4$ (or $C_4$ to $C_3$) is likely preferentially selected by nature. Therefore, violations of the assumption for maximum parsimony method exists, and the result is prone to bias. But a rough estimate could still help with timing the origins of $C_4$ and show the distribution of $C_4$ among the lineages of interest. For my results, incongruence among different analyses will be discussed (this is mainly result from different topologies). Since the basal subfamilies and BOP clade are all unambiguously $C_3$, the ancestral state of the whole Poaceae family is presumably $C_3$.

### 2.1.4 Gene family analysis of *ppc*

There are more than 20 genes reported to be involved in $C_4$ photosynthesis (Christin et al., 2015; Moreno-Villena et al., 2018; Sedelnikova et al., 2018), and not all of them has been studied systematically. These genes are either important for biochemical pathways or anatomical structures of $C_4$ photosynthesis. In my project I'll start gene family analyses from the *ppc* genes that encode for the enzyme phosphoenolpyruvate carboxylase (PEPC) which functions in the initial step of $CO_2$ fixation in all Poaceae $C_4$ lineages as well as in many other plant families.

The objective is to identify and classify all sampled $C_4$ and non-$C_4$ *ppc* sequences in this project into the six gene clades described in previous studies (*ppc-aL2, ppc-aL1a, ppc-aL1b, ppc-aR, ppc-B1* and *ppc-B2*; Christin and Besnard, 2009) based on gene family tree and sequence features. The topology of the gene family tree will be checked for potential gene duplication and gene loss events. In addition, potential horizontal gene transfer (HGT) events could be identified. I plan to include representative species from each Poaceae subfamily, including transcriptomes and genomes sampled in this project, as well as *ppc* sequences from public data sets when necessary. Species from other Poales will be used as outgroups. The homologs of *ppc* sequences will be retrieved from transcriptomes/genomes by blast search, and sequence below a length threshold will be filtered. A gene family tree of hundreds of gene transcripts will be reconstructed, and *ppc* gene clades will be identified using previously reported sequences as references. In cases where the gene tree topology (within a gene clade) differs significantly from species phylogeny, gene duplications/losses or horizontal gene transfer or could be proposed. Potential $C_4$ type *ppc* genes will be identified by checking the corresponding amino acid sequences, especially #780 (maize numbering, NP_001154820.1) which was reported to be critical for $C_4$-specific enzymatic characteristics (Blasing et al., 2000).

One shortcoming of using transcriptome data is that not all *ppc* genes are guaranteed to be sampled; *ppc* transcripts could be missing due to lower expression in the tissues we sampled. To test the feasibility of using transcriptome data for the gene family analyses of *ppc*, I will first select a subset of species covering all subfamilies and reconstruct the gene tree using *ppc* homologs from these species. If most ppc genes could be obtained by transcriptome data, I will expand the analysis to include more species.

I anticipate categorizing all sampled *ppc* genes into the six *ppc* gene clades and identifying $C_4$ type *ppc* genes. Gene family analyses will be able to reveal duplication/loss patterns of *ppc* genes. In addition, detailed comparison of the gene sequences and corresponding amino acid sequences could identify critical sites under natural selection.

## 2.2 Methods

### 2.2.1 Grass sample collection and sequencing

Taxon sampling in this project aimed to represent Poaceae with all subfamilies and as many tribes as possible. For large tribes (for example, Andropogoneae in Panicoideae, Cynodonteae in Chloridoideae, and Poeae in Pooideae), we also tried to include as many subtribes as possible. We sampled a total of 357 Poaceae species, representing 45 of 52 tribes in Poaceae. In addition, we sampled thirteen outgroup species, including one species from each of Ecdeiocoleaceae (*Ecdeiocolea monostachya*) and Joinvilleaceae (*Joinvillea ascendens*), which together form a sister clade to Poaceae; also sampled are members of other Poales families, including Flagellariaceae, Restionaceae, Eriocaulaceae, Cyperaceae, Juncaceae, and Typhaceae, as well as those in three other orders close to Poales: Arecales (Arecaceae), Zingiberales (Zingiberaceae), and Commelinales (Commelinaceae). Information of taxa included in this project is listed in Supplemental Table 1 of Huang et al. (2022).

Fresh plant samples were collected from the field and preserved in paper bags filled with silica-gel to keep dry; or brought back to lab shortly followed by RNA extraction. For those species with no fresh material available, herbarium samples were carefully collected from specimen and also kept in paper bags. Total RNA/DNA was isolated from samples of leaves, stems, inflorescences, or young fruits using the RNA isolation kit NucleoSpin RNA Plant (REF 740949.50) by MACHEREY-NAGEL or by standard EDTA protocol. The RNA/DNA samples were used for library construction and sequencing by either the Penn State core facility or commercial sequencing companies, using the Illumina platform to construct sequencing libraries and perform pair-end sequencing to obtain 150-bp reads. The procedure generally include the following steps: (1) total RNA was extracted from fresh, frozen or dried plant tissues, then treated with DNase to remove

DNA; (2) mRNAs were captured by purification using a column with oligo (dT); (3) mRNAs were used as templates to synthesize first-strand cDNA using random hex-mer primers; (4) second-strand cDNAs were synthesized and purified, their 5' ends repaired and 3' end adenylated, finally ligated to adaptors; (5) cDNAs were amplified by PCR. Paired-end transcriptome sequencing (2x150 bp) was done by GENERGY BIO using Illumina Hiseq3000. Genome sequencing (2x150 bp) was done by GENERGY BIO using Illumina Hiseq3000. Public transcriptomes/genomes/SRA data were retrieved from NCBI databases (https://www.ncbi.nlm.nih.gov/) and EMBL-EBI (http://www.ebi.ac.uk/). See Supplemental Table 1 of Huang et al. (2022) for the sources of samples.

### 2.2.2 Sequence trimming and assembly

The procedures for transcriptomic and genomic sequence processing and assembly are illustrated in Figure 2-1. For transcriptomic data, paired sequencing data sets were first trimmed by Trimmomatic (as Trinity 2.2.0 plug-in; Grabherr et al., 2011) using default settings. FastQC (0.11.8) (Andrews et al., 2015) quality checks were performed after trimming to confirm the removal of adapters and low-quality regions. Transcriptome assembly was performed using Trinity (V 2.2.0) with default parameters on the Penn State ACI server. De-duplication of assembly contigs was done by CD-HIT-EST (V 4.6.8) (Fu et al., 2012) with the parameter -c 0.98. Coding sequences were extracted from deduplicated contigs by TransDecoder (V 5.3.0) (http://transdecoder.sourceforge.net). For shotgun genome sequencing data, Trimmomatic was also implemented to remove sequencing adaptors and low-quality regions. Kmergenie (1.7048) (Chikhi and Medvedev, 2014) was used to optimize the value of k-mer in subsequent assembly process (only used for Puelioideae samples). Optimized K values were set for the assembly of genomic data sets by SOAPdenovo2 (Luo et al., 2012). Assembled genomic contig data sets were deduplicated by CD-HIT-EST (V 4.6.8). Considering that: (1) the genomic data generated by shotgun genome

sequencing were relatively sparse, and some target coding sequences may be partial; (2) genomic sequences contain introns and other elements that cannot be readily identified, to obtain more sequence for subsequent analyses, the assembled genomic contigs were not processed by Transdecoder to generate cds data sets. For public genomes/transcriptomes, non-redundant coding sequences were retrieved directly from NCBI database. SRA data sets were processed same as the transcriptome data sets generated by our own project. Statistics on none-redundant coding sequence data sets and genomic contig data sets were calculated by statswrapper.sh (a bbmap tool, V 38.33) to check assembly quality and are provided in Supplemental Table 1 in Huang et al (2022).

### 2.2.3 Selecting and obtaining target low-copy orthologous nuclear genes from transcriptomic and genomic data

The genome/transcriptome sequences of ten Poaceae species (*Brachypodium distachyon, Eleusine coracana, Hordeum vulgare, Lygeum spartum, Oryza sativa, Phaenosperma globosum, Phyllostachys heterocycle, Setaria italica, Sorghum bicolor, Stipa aliena*) that represent five largest subfamilies but avoiding the recent polyploids wheat and maize were selected, with additional criteria on data quality, to identify putative low-copy (one or two copies per species) nuclear genes across Poaceae. Such putative orthologous genes were identified by using OrthoMCL v1.4 (Li et al., 2003), with the following parameters: perl orthomcl.pl --mode 3 --blast_file 10sps.blastresult --gg_file 10sps.gg, where "--mode 3" instructs OrthoMCL to perform analysis using user-provided BLAST output file (10sps.blastresult) and the genome gene relation file (10sps.gg), and additional default settings. The HMM files of 1,234 OGs (orthologous groups, or genes) identified were used as the seeds for HaMStR (13.2.6) (Ebersberger et al., 2009) to search and retrieve the corresponding orthologous sequences from the assembled contig datasets from transcriptome and genome sequencing. Cutoff e-values for blast and hmm search were both set to 1e-20. There is only one

sequence retained per dataset for each seed, and sometimes fragments matching non-overlapping parts of the seeds were combined to represent the whole sequence. The number of orthologous sequences retrieved for each genome contig (sampled in this project) dataset ranges from 252 to 1018, and the number of orthologous sequences retrieved for each cds dataset (all others except for those seven genome-skimming datasets) ranges from 235 to 1,234.

### 2.2.4 Sequence alignment and reconstruction of single-gene Maximum Likelihood trees

Retrieved orthologous sequences by HaMStR were sorted by sequence ID (orthologous group ID, or gene ID) and reorganized and formatted into fasta format files. Nucleotide sequences of each OG were aligned by MAFFT (v7.397) (Katoh et al., 2009) using the --auto option. Alignments were then trimmed by trimAl (1.4.1) (Capella-Gutiérrez et al., 2009) with -automated1 option to remove poorly aligned regions and/or sequences. Single-gene ML trees on the alignments of 1,234 OGs were reconstructed by RAxML (8.2.1) (Stamatakis, 2014) with rapid bootstrapping of 100 replicates and GTRCAT model.

### 2.2.5 Detection of sequences prone to long branch attraction

To identify and remove genes that are prone to exhibit long branch attraction, TreSpEx (1.1) (Struck, 2014) was applied on all the 1,234 single-gene alignments together with single-gene ML trees corresponding to orthologous genes to analyze long-branch attraction (determined by heterogeneity or longest branches) and saturation (determined by slope or R2 of linear regression). The probability density function curves of these four indicators are plotted by R (Supplemental Figure 2 of Huang et al., 2022). Genes that are deviated from normal distribution by each of the four indicators were removed. The numbers of genes removed are 389 and 393 due to heterogeneity

or longest branches, respectively; in addition, 555 and 96 genes were removed according to slope or R2 of linear regression, respectively. After deletion of the genes from these four sets, 571 genes out of the 1150-gene set were retained and was further filtered for super-matrix and molecular clock analysis.

### 2.2.6 Phylogenetic analyses using the Astral coalescent method or a supermatrix dataset and au test

For the coalescent analysis dataset with 378 samples, the number of samples with a positive detection for each gene ranges from 23 to 377. Since the number of genes retrieved by searches using HaMStR from the six shot-gun genomic contig datasets are relatively low (the lowest being 252), and the six genomic datasets represent the two genera, *Guaduella* and *Puelia* from one of the basal subfamily Puelioideae, we filtered the set of genes to make sure that each gene is present in at least one species from each of the two genera. The remaining 1,150 genes were further filtered by coverage and alignment length to generate smaller sets. We have six sets of genes for the coalescent analyses, their sizes (number of genes) are 1,150, 895, 775, 570, 436 and 180, respectively. See Figure 2-1 for the procedure of gene selection.

Astral 5.6.3 (Sayyari and Mirarab, 2016) was used to infer multi-gene coalescent trees from different sets of single-gene trees. There are two major methods to evaluate the support of coalescent results, namely multi-locus bootstrapping (MLBS) and local posterior probability (PP). MLBS method is prone to biased estimate of the support and thus affect interpretation of phylogenetic trees (Wickett et al., 2014; Bayzid et al., 2015). On the other hand, Bayesian posterior probabilities as a support for phylogenetic tree topology are shown to be more precise and reproducible (Sayyari and Mirarab, 2016). Therefore, in this project we preferred PP values as the support for phylogenetic trees. The coalescent trees are edited by Dendroscope (V 3.6.2) (Huson

and Scornavacca, 2012) and summarized by TreeGraph2 (2.14.0) (Stöver and Müller, 2010). The 180-gene set was also used to generate a supermatrix dataset with a length of 184,993 and a total of 71,037,312 matrix cells. The percentage of missing sites is 10.977%, and the proportion of variable sites is 0.798. AU (approximately unbiased) test was performed using CONSEL v0.20 (Shimodaira and Hasegawa, 2001) on the 180-gene super-matrix with sequences from 384 species (Supplemental Figure 5 in Huang et al., 2022).

### 2.2.7 Reconstruction of ancestral state for photosynthetic pathway

The reconstruction of ancestral state of photosynthetic pathway type was performed by Mesquite (3.6) (Massidon and Maddison, 2019). The state of sampled species is coded as either 0 (C3) or 1 (C4) according to information summarized by Soreng et al. (2017), and the ancestral state is inferred by Most Parsimony method using default parameters in the context of the topology from five coalescent trees, respectively (see Supplemental Table 1 in Huang et al., 2022 for the state code of photosynthetic type).

### 2.2.8 Estimating divergence time of Poaceae lineages

In our analyses a total of thirteen fossil calibrations are used, including phytolith data for Poaceae (see Appendix for fossils used), which provide informative calibration points and support older ages than those estimated using only the relative scarce macrofossils (Christin et al., 2014; Kellogg, 2015). The phytoliths from grasses (silica bodies) are regarded as distinctive from other families of Poales and can be assigned to subfamilies of Poaceae (Magallón et al., 2015). Taxonomic assignment and age of the fossils are designated according to references cited in

Supplemental Table 6. In our analyses, all the fossil calibrations were implemented as minimum constraints except for the crown age of Commelinids, which is set to be no older than 118 Mya.

Considering the large amount of sequence data from over 380 taxa, we used PL method implemented in treePL (1.0) (Smith and O'Meara, 2012) to estimate the divergence time. The ML tree reconstructed by RAxML (8.2.1) from the smallest set of 180 genes with branch length information was used as the input tree, to avoid likely systematic errors from supermatrix datasets of hundreds of genes (Philippe et al., 2011). This tree was generated using the 180-gene set and with the topology of the 1,150-gene coalescent tree (also supported by most analyses) as a constraint. Parameter optimization and cross-validation were done to decide the best smoothing value along with other parameters. The smooth value is decided as 0.1, which is low and indicates a large deviation from the strict molecular clock hypothesis (Huang et al., 2016a). One hundred BS replicates with branch length information of the 180-gene ML tree were also generated by RAxML (8.2.1) to calculate the confidence intervals of node ages.

### 2.2.9 *ppc* gene family analysis

The sampling for *ppc* gene family analysis was done as to represent all the subfamilies in the PACMAD clade and to cover most C4 lineages. In addition, species from other subfamilies of Poaceae were included to cover major tribes, excluding subfamily Puelioideae, which we only have genome skimming data available and it is shown that *ppc* homologs cannot be reliably retrieved by blast search. Fifteen species from nine other families of Poales, as well as Musaceae (Zingiberales) and Asparagaceae (Asparagales) were also included as outgroups. A total of 107 samples were included for the ppc gene family analysis. Amino acid sequences representing the six ppc lineages from *Cyrtococcum patens* (Panicoideae) were used as queries to perform tblastn against the coding sequence datasets of the selected species. The six reference coding sequences from *Cyrtococcum*

*patens* and some other species from public data sets were also included (see Supplemental Table 8). Duplicate copies with identical sequences from the same samples were removed, and coding sequences significantly shorter than others were also removed manually, but critical C4 type *ppc* sequences (sequences that belong to species that are in critical phylogenetic positions) were kept. A total of 516 ppc sequences were retained. The *ppc* sequences were translated into amino acid sequences and aligned by ClustalO (1.2.4) (Sievers et al., 2011). Alignment of nucleotide sequences was then generated based on the corresponding amino acid alignment and was trimmed by trimal (1.4.1) (Capella-Gutiérrez et al., 2009) to remove poorly aligned regions. ML analysis was conducted on the trimmed alignment by iqtree (1.6.12) (Nguyen et al., 2015) to reconstruct gene family trees. C4 *ppc* genes were distinguished by the serine at position 780 (following the numbering in *Zea mays* C4 *ppc*, accession number: GRMZM2G083841) in the corresponding amino acid sequences (Bläsing et al., 2000). Sequences with other amino acids at this residue in the alignment were not treated as C4 ppc genes, as such sequences have not been shown to be C4 *ppc* genes experimentally.

## 2.3 Results

### 2.3.1 Generation of new transcriptomic and genomic datasets and selection of nuclear genes

For nuclear phylogenetic analyses, we sequenced 342 transcriptomes and seven genomes (by genome skimming) with a median number of 68,153 unigene sequences and an average N50 value of 934 bp (see Supplemental Table 1 in Huang et al., 2022 for more statistics per data set). These and 35 public datasets represent 371 Poaceae samples (Anomochlooideae 2; Aristidoideae 4; Arundinoideae 5; Bambusoideae 51; Chloridoideae 86; Danthonioideae 7; Micrairoideae 3; Oryzoideae 16; Panicoideae 79; Pharoideae 1; Pooideae 111; Puelioideae 6; 14 samples were redundant) and 13 outgroups. The taxon sampling here includes 45 of the 52 tribes, whereas the remaining 7 un-sampled tribes have a total of ~40 species.

We used genomic/transcriptomic sequences of ten Poaceae species from large subfamilies (see Figure 2-1) to identify 1,234 conserved low-copy nuclear genes and searched for their homologous sequences from all other datasets. Because the six species of Puelioideae (with three species in *Guaduella* and three in *Puelia*) had genome skimming datasets with relatively shallow sequencing depth, we maximized the gene coverage of Puelioideae by selecting for genes that have homologs in at least one species in each of *Guaduella* and *Puelia*, resulting in 1,150 genes. To reach relatively high taxon coverage, we selected for genes with at least 90% coverage among sampled taxa, yielding 895 genes. The coalescent method for phylogenetic reconstruction uses single gene trees; thus, to ensure the quality of each gene tree, we favored longer genes with more phylogenetic information to obtain gene trees with relatively high support values. We selected three additional sets of 775, 570, and 436 genes with progressively longer cutoffs of the alignment length (see Figure 2-1 for a workflow). Furthermore, we examined the original 1,234 set and removed genes that might be more prone to long branch attraction to generate a set of 579 genes; the overlap

of these with the previously identified 1,150 genes, plus an additional coverage requirement of at least 370 taxa, resulted in a set of 180 genes. This smallest gene set was used for phylogenetic analysis using the Maximum Likelihood method with a super-matrix approach because of the known systematic errors when super-matrix datasets with large gene sets are used in phylogenetic reconstructions (Philippe et al., 2011).

### 2.3.2 A Highly supported Poaceae phylogeny – early divergent lineages

For a nuclear Poaceae phylogeny, we used the 1150, 895, 775, 570, and 436 gene sets for coalescent analyses (summarized in Figures 2-5, 2-6, 2-7 and 2-8). Our results are consistent, with maximum local posterior probability values on most branches [321/ (364-1), 88.43%], and agree with accepted classifications for most taxonomic groups from subfamily to genus levels. Phylogenies using supermatrix method with the 180-gene set were also generated.

The results from all analyses support the monophyly of 11 subfamilies but not Puelioideae, which is divided into two highly supported paraphyletic branches, one for each of *Guaduella* and *Puelia* (Figure 2-5). Previously, the monophyly of Puelioideae was supported using three genes from two species, *Puelia ciliata* and *Guaduella marantifolia* (Clark et al., 2000). Other Poaceae phylogenetic studies that included Puelioideae sampled one species, *Puelia olyriformis*, and supported the placement of Puelioideae as the third divergent subfamily before the separation of the BOP and PACMAD clades (GPWG II, 2012; Jones et al., 2014; Saarela et al., 2018). Although the monophyly of Puelioideae could not be rejected by approximately unbiased (AU) tests with the 180-gene supermatrix dataset, the difference in monophyly of Puelioideae between this study and previous results could be explained by different history of nuclear and plastid genes. Consistent with all previous studies, Anomochlooideae is always the first diverging lineage, followed by Pharoideae and the two genera of Puelioideae (Figure 2-5).

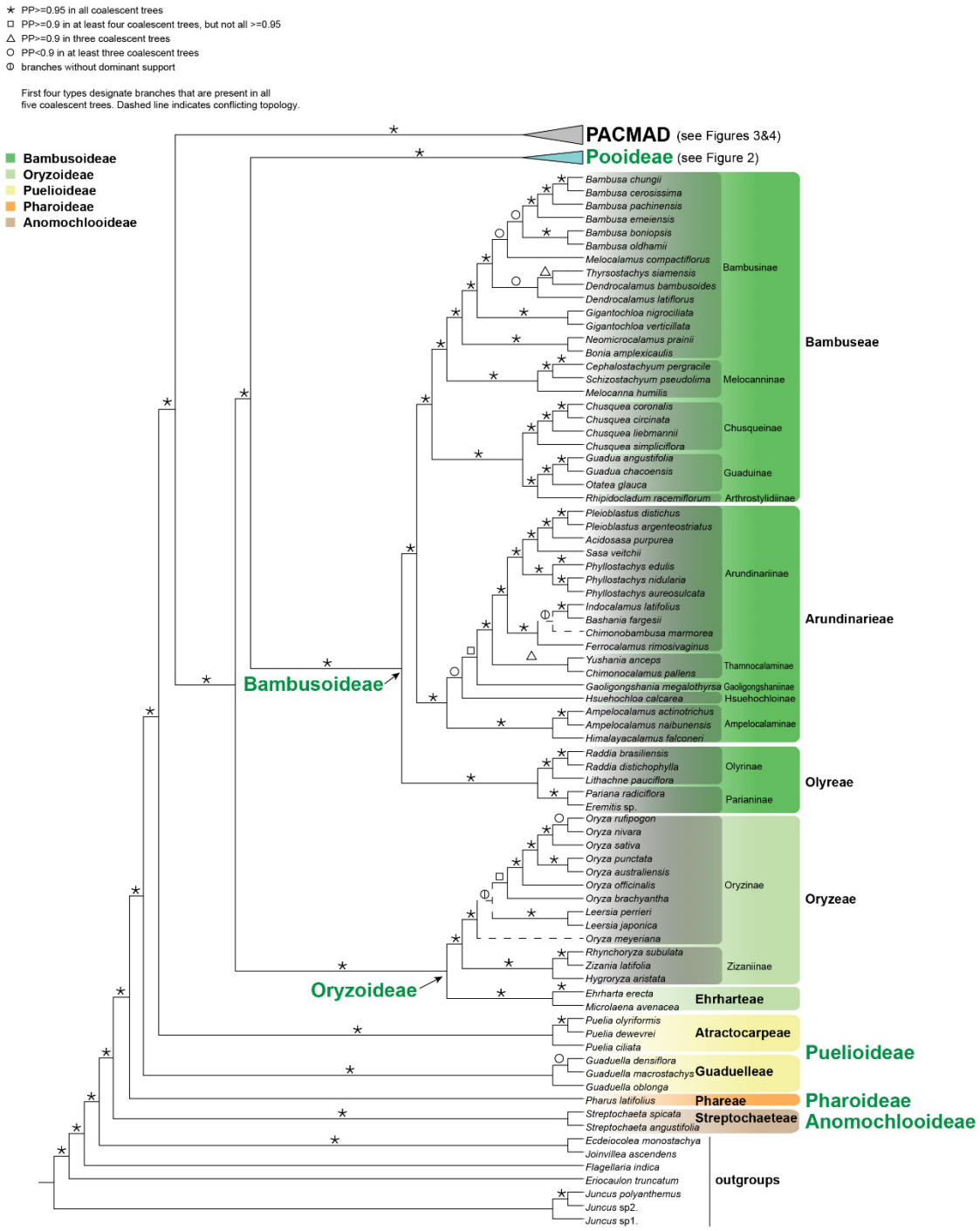


Figure 2-5: A summary for a portion of the Poaceae phylogeny (including Bambusoideae and Oryzoideae). A portion of the Poaceae phylogeny is shown here for a summary of results from coalescent analyses using five different gene sets (with 1150, 895, 775, 570, and 436 genes); detailed phylogenetic relationships are shown for species in the three small, early divergent subfamilies (Anomochlooideae, Pharoideae, Puelioideae), Bambusoideae and Oryzoideae in the BOP clade. Symbols above the branches represent local PPs, and the corresponding values are indicated in the upper left corner.

Pooideae and the PACMAD clade are represented by collapsed triangles, and the detailed phylogenetic relationships for these clades are shown in Figures 2-5, 2-6, and 2-7. Different colored backgrounds represent subfamilies, as explained in the upper left corner. Names of subfamilies are shown in green, and names of tribes are shown in black. Branches associated with alternative topologies are shown in dashed lines. The symbols for PP values are follow the same rules in Figures 2-5, 2-6 and 2-7.

### 2.3.3 Phylogenetic relationships in the BOP clade

The BOP clade with Bambusoideae, Oryzoideae, and Pooideae was first identified by Clark et al. (1995) and is monophyletic in several studies with alternative relationships of the three subfamilies; however, the topology (O, (B, P)) is supported by recent studies using plastid genes or whole plastomes (GPWG II, 2012; Zhao et al., 2013; Jones et al., 2014; Saarela et al., 2018). The same topology is supported maximally and consistently by our results (Figure 2-5).

In Oryzoideae, the two tribes sampled here, Ehrharteae and Oryzeae, are monophyletic, as are two Oryzeae subtribes, Oryzinae and Zizaniinae (Figure 2-5). In Oryzinae, a clade of seven sampled *Oryza* species with *O. sativa* (subspecies japonica of the cultivated rice) and two closely related species *O. nivara* and *O. rufipogon*, received moderate support, consistent with the close but complex relationships among these three species (Zhu and Ge, 2005). In addition, the relationships of other *Oryza* species relative to *O. sativa* and *O. rufipogon* are different from previously reported relationships (Kellogg, 2009; Tang et al., 2010a). *Oryza meyeriana* and *O. granulata* have been considered as the same species previously (Ge et al., 1999); *O. meyeriana* is placed at two different positions in different coalescent results, and its position was uncertain in previous studies (Aggarwal et al., 1999; Ge et al., 1999; Zou et al., 2008; Kumagai et al., 2010; Zou et al., 2013). In Zizaniinae, *Hygroryza* is sister to *Rhynchoryza* + *Zizania*, consistent with previous studies (Kellogg, 2009; Tang et al., 2010a).

In Bambusoideae, the three tribes Olyreae, Arundinarieae and Bambuseae are each revealed to be monophyletic with maximal support (Figure 2-5), with a topology of (O, (A, B)). The sister relationship of the woody tribes Arundinarieae and Bambuseae supports a single origin

of woodiness in bamboos. Previously, a plastome phylogeny placed Arundinarieae as sister to the other bamboos (Wysocki et al., 2015), supporting two origins or one loss of woodiness in Olyreae. A phylogeny of 38 bamboo species (Triplett et al., 2014) and a recent genome-based analysis (Guo et al., 2019) both strongly supported the herbaceous Olyreae being sister to the woody bamboos.

In Olyreae, two subtribes are monophyletic with maximum support in all trees. Members of Arundinarieae are temperate woody bamboos and were previously placed in a single subtribe Arundinariinae but are recently divided into five subtribes (Zhang et al., 2018; Zhang et al., 2020; Guo et al., 2021). The phylogeny is consistent in all coalescent trees, except the position of *Chimonobambusa marmorea* (in Arundinariinae). Among the five subtribes, Ampelocalaminae is placed as sister to the remaining four subtribes, with Hsuehochlinae and Gaoligongshaniinae consistently being successive sisters of the combined clade of Arundinariinae and Thamnocalamaminae.

In Bambuseae with tropical woody bamboos, 5 of 11 subtribes are sampled; Guaduinae and Arthrostylidiinae are sisters and together with Chusqueinae form a neotropical clade, consistent with previous studies (Wysocki et al., 2015). On the other hand, Melocanninae and Bambusinae form a paleotropical clade. In Bambusinae, ten sampled species belong to the *Bambusa*-*Dendrocalamus*-*Gigantochloa* complex (Goh et al., 2013), where *Gigantochloa* is monophyletic and sister to a highly supported clade of the other two genera. Bambuseae and Arundinarieae were reported to have allopolyploid ancestry, with Arundinarieae being tetraploids with subgenomes designated A and B, while Bambuseae includes both tetraploids (neotropical; subgenomes C and D) and hexaploids (paleotropical; subgenomes C, D and E) (Triplett et al., 2014). A recent study of diploid and polyploid woody bamboo genomes presented an alternative hypothesis with ABCD sub-genomes, including subgenome C shared by the three woody bamboo lineages (Guo et al., 2019). Such polyploid history of woody bamboos suggest that their phylogenetic relationships are probably more complex than presented here (see next section for more discussion).

Pooideae is the largest among all 12 Poaceae subfamilies and includes wheat, barley and other crops, as well as the model grass *Brachypodium distachyon* (Vogel et al., 2010). My sampling covered 111 samples in 15 tribes and results maximally support the monophyly for seven out of eight tribes with at least two species (in the order from early to late divergent lineages): Duthieeae, Meliceae, Stipeae, Brachypodieae, Poeae, Bromeae, and Triticeae (Figure 2-6). However, Diarrheneae with two species *Diarrhena obovata* and *Neomolinia japonica*, is polyphyletic. Ten of the 15 Pooideae tribes are grouped into five supertribes (recognized by Soreng et al., 2017), which are all maximally supported as monophyletic. The separation of the monotypic Phaenospermateae from Duthieeae is consistent with recent reports by Schneider et al. (2011). The supertribe Stipodae with tribes Stipeae and Ampelodesmeae is not monophyletic, as *Diarrhena* of Diarrheneae is embedded and sister to Stipeae. The clade with Stipeae, *Diarrhena*, and Ampelodesmeae is sister to a large clade with five tribes plus *Neomolinia* of Diarrheneae. In this clade, Neomolinia is clearly the first to diverge, followed by Brachypodieae. Poeae is sister to Littledaleeae, Bromeae and Triticeae combined.

### 2.3.4 Phylogenetic relationships in the PACMAD clade

The PACMAD clade as a whole and its six subfamilies are maximally supported to be monophyletic in a number of studies, although the relationship among the subfamilies is inconclusive and sensitive to phylogenetic methods (GPWG II, 2012; Cotton et al., 2015; Soreng et al., 2017; Saarela et al., 2018). Nevertheless, increasing evidence supports Aristidoideae as sister to the other five subfamilies (Vicentini et al., 2008; GPWG II., 2012; Kellogg, 2015; Soreng et al., 2015; Soreng et al., 2017). In all the five coalescent analyses of my project, Aristidoideae is sister to the other five subfamilies and consistently received maximum support (Figure 2-7).

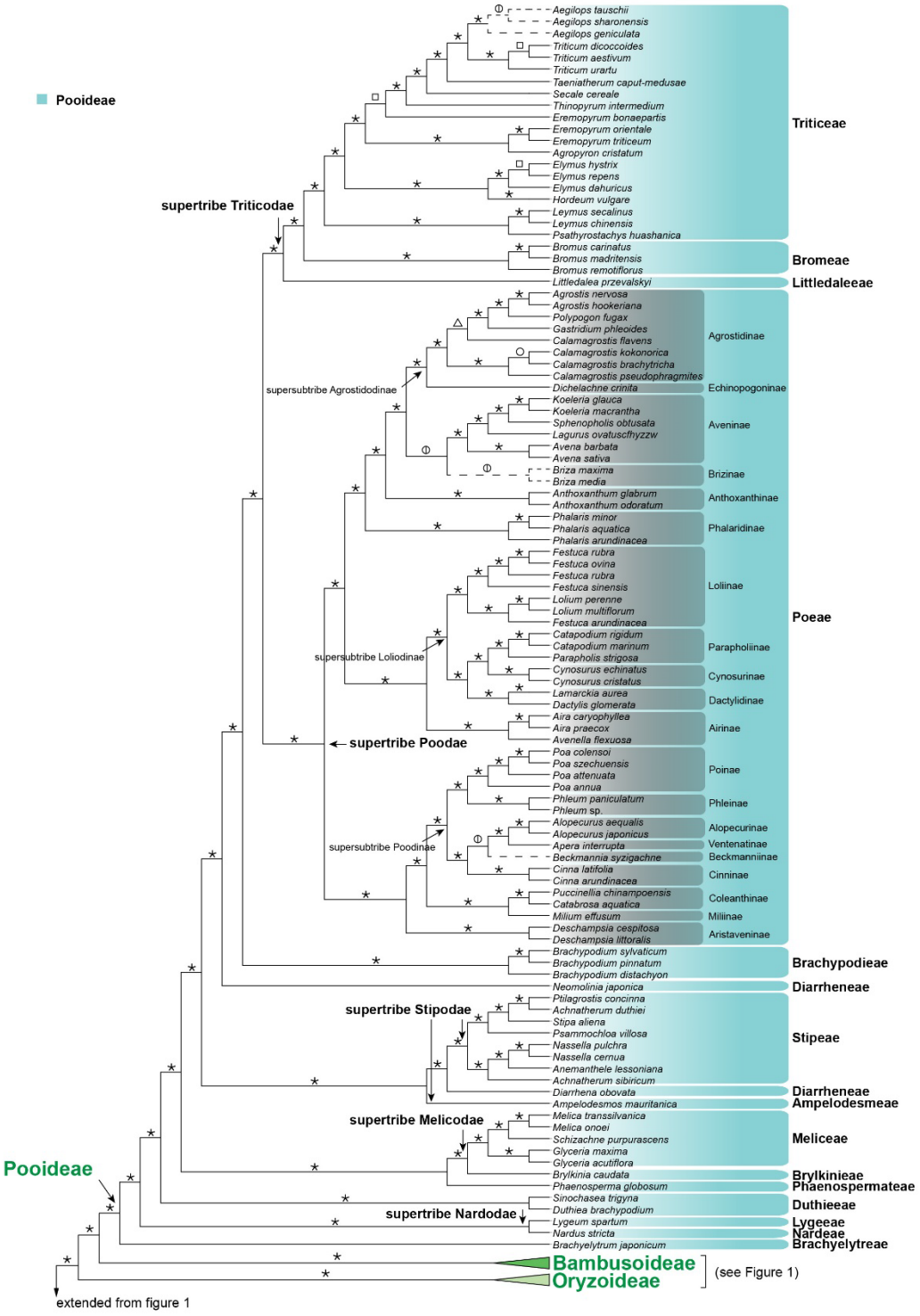


Figure 2-6: A summary for a portion of the Poaceae phylogeny (Pooideae). The Pooideae portion of the summarized Poaceae phylogeny is shown. Supertribes are indicated with arrows pointing to the nodes, as are three supersubtribes in the tribe Poeae. See also legend for Figure 2-5.

The relationships among four of the remaining five subfamilies are also consistent and highly supported as (((Chloridoideae, Danthonioideae), Arundinoideae), Panicoideae), but not the position of Micraroideae, which is placed at one of two positions with varying support values (see next section for more discussion).

My results also provide strong support for many relationships within the PACMAD subfamilies, with the relationships among three genera of Aristidoideae being consistent with the GPWG II phylogeny (GPWG II., 2012). In Panicoideae (~3,240 species), the largest subfamily in PACMAD, my sampling included ten tribes and the six tribes with two or more species are all monophyletic (Figure 2-7). The tribes Chasmanthieae and Zeugiteae form a sister clade to the remaining tribes of Panicoideae, with Centotheceae and Thysanolaeneae in the next divergent clade, followed successively by Gynerieae and Tristachyideae. As a comparison, previous studies (Sánchez-Ken et al., 2007; GPWG II, 2012; Saarela et al., 2018) supported a branch with three tribes (Tristachyideae + (Thysanolaeneae + Centotheceae)) as sister to all other Panicoideae tribes and (Chasmanthieae + Zeugiteae) on the next divergent branch.

The remaining four Panicoideae tribes belong to two maximally supported monophyletic supertribes, Panicodae and Andropogonodae. Panicodae has only one tribe, Paniceae, with six out of seven subtribes sampled, and *Sacciolepis indica,* which was not previously assigned to a subtribe. The subtribes are monophyletic except for Panicinae (*Panicum*) and have consistent relationships (except for placement of Dichantheliinae), but the support for monophyly of Boivinellinae is lower than those for other subtribes and other topologies are possible. *Panicum brevifolium* is maximally supported as sister to *Sacciolepis indica*, apart from other *Panicum* species. Also, *Pennisetum* is nested in a clade with *Cenchrus* species, consistent with the recent treatment of *Pennisetum* as a synonym of *Cenchrus* (Chemisquy et al., 2010). Two *Setaria* species are not grouped together, with *Setaria palmifolia* next to the clade of *Cenchrus/Pennisetum* and *Setaria italica* being sister to

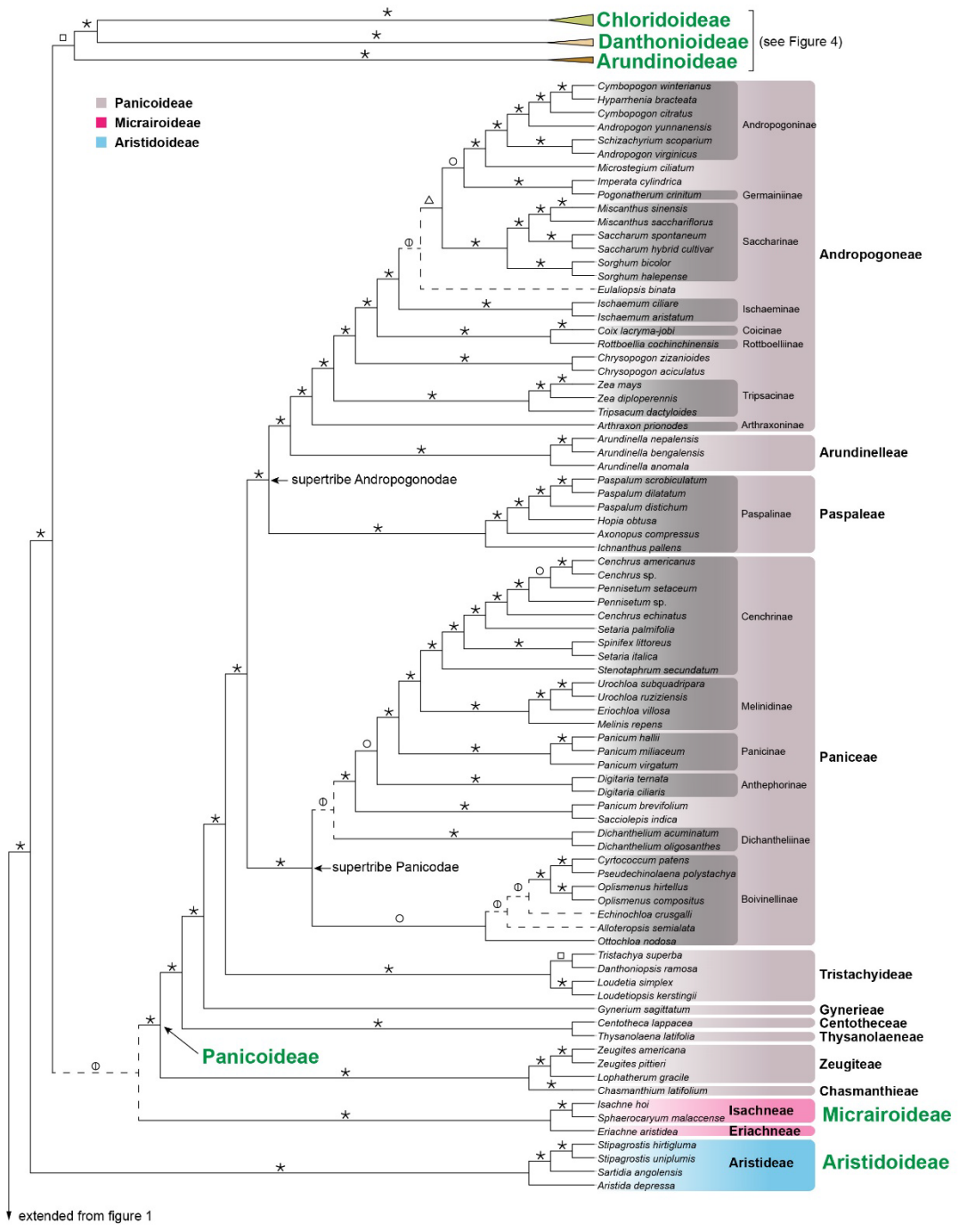


Figure 2-7: A summary for a portion of the Poaceae phylogeny (Aristidoideae, Micrairoideae, and Panicoideae). A portion of the summarized Poaceae phylogeny is shown, with three subfamilies, Aristidoideae, Micrairoideae, and Panicoideae, all belong to the PACMAD clade. Supertribes are marked with arrows pointing to the nodes. See also legend for Figure 2-5.

*Spinifex littoreus,* consistent with previous studies showing that *Setaria* is not monophyletic and the placement of *Setaria palmifolia* and *Setaria italica* in separate lineages (Morrone et al., 2012).

The other supertribe in Panicoideae, Andropogonodae, has three previously defined tribes, Paspaleae, Arundinelleae and Andropogoneae, and a recently described tribe Jansenelleae (Bianconi et al., 2020), which includes two genera not sampled here. The three sampled tribes are maximally supported as monophyletic, with Paspaleae being sister to Arundinelleae plus Andropogoneae, consistent with previous reports (GPWG II., 2012; Saarela et al., 2018). In Paspaleae, *Ichnanthus*, *Axonopus*, and *Hopia* consistently form a grade in all trees here outside a clade of three *Paspalum* species. In Andropogoneae, our sampling includes eight subtribes and four unplaced genera, *Chrysopogon*, *Eulaliopsis*, *Imperata,* and *Microstegium*. Our analyses support the monophyly of subtribes Tripsacinae, Ischaeminae, and Andropogoninae, but not Saccharinae. In addition, the placement of Arthraxoninae and Tripsacinae as successive sisters to other Andropogoneae is consistent with a previous study using plastomes (Saarela et al., 2018), but not with the topology in another nuclear phylogeny (Estep et al., 2014). The next lineage to diverge has two *Chrysopogon* species, supporting a recently proposed designation of this genus as a new subtribe (Welker et al., 2020). The subtribes Rottboelliinae and Coicinae form a clade sister to the remaining Andropogoneae with four subtribes, which were not resolved previously (Mathews et al., 2002). The previously unplaced *Eulaliopsis* is either sister to a clade with the subtribes Saccharinae, Germainiinae and Andropogoninae, or placed elsewhere, while *Microstegium* is maximally supported as sister to Andropogoninae.

Arundinoideae is represented here (Figure 2-8) by four genera/species belonging to two tribes, Arundineae and Molinieae, the latter of which has two subtribes Crinipinae and Molininae. The placement of *Pratochloa walteri* in Crinipinae is in agreement with a previous study (Ingram et al., 2011). In Danthonioideae (one tribe Danthonieae), *Danthonia* is monophyletic, but *Cortaderia* is not, in agreement with the reported paraphyly of *Cortaderia* (Barker et al., 2003).

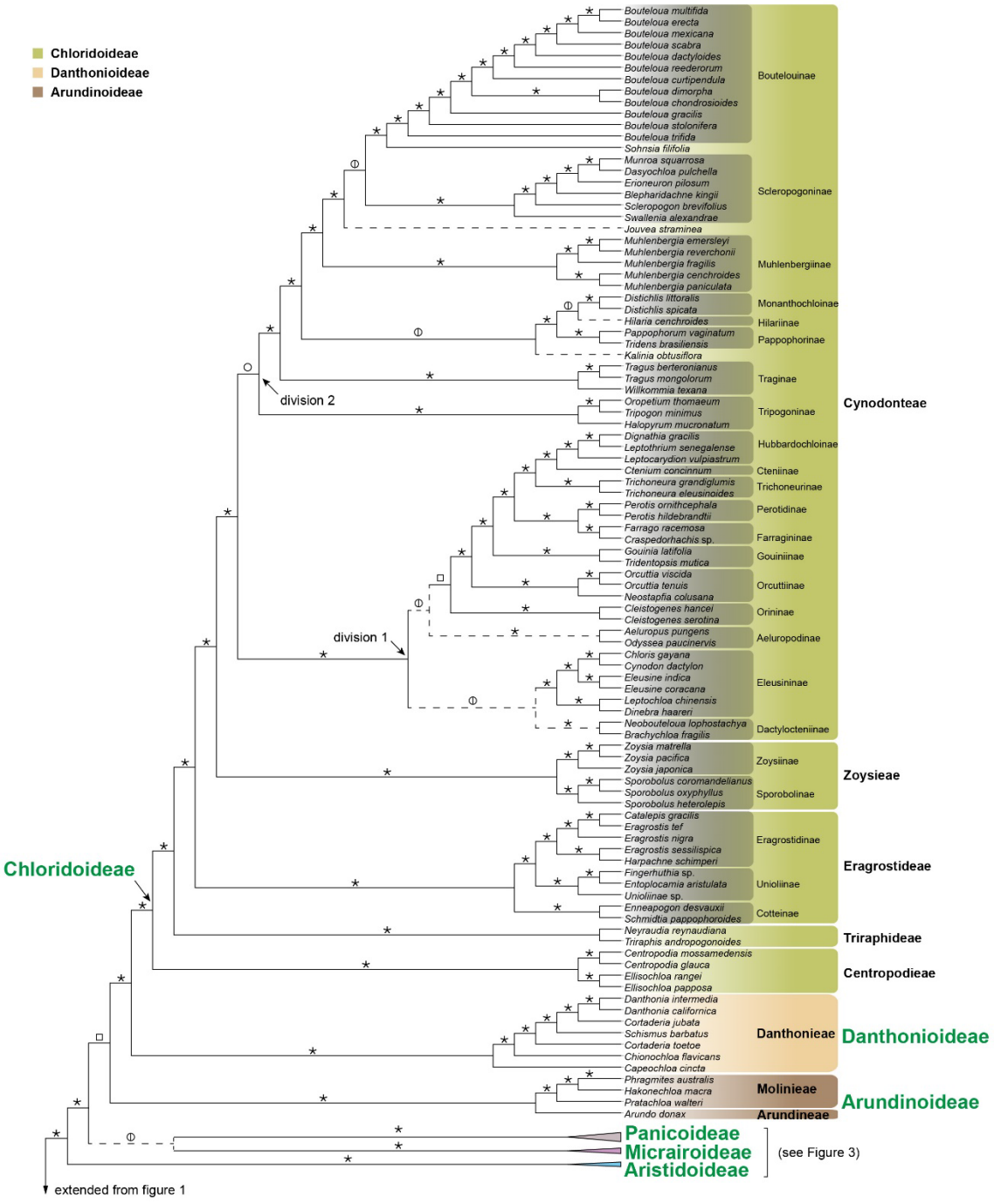


Figure 2-8: A summary for a portion of the Poaceae phylogeny (Arundinoideae, Danthonioideae, and Chloridoideae). A portion of the summarized Poaceae phylogeny is shown, with three subfamilies, Arundinoideae, Danthonioideae, and Chloridoideae, all belong to the PACMAD clade. Two large clades in the large tribe Cynodonteae are marked with "division 1" and "division 2". See also legend for Figure 1.

Chloridoideae (~1,600 species) is the second largest subfamily in PACMAD, and my study included 86 samples in 56 (out of 124) genera; all five tribes, Centropodieae, Triraphideae, Eragrostideae, Zoysieae, and Cynodonteae, are maximally supported as monophyletic (Figure 2-8). Centropodieae, with two monophyletic genera, is maximally supported as sister to other Chloridoideae tribes. Triraphideae, Eragrostideae and Zoysieae are monophyletic and occupy the next three successive sister lineages. Within Eragrostideae, all three subtribes are monophyletic, with Cotteinae being sister to a clade of Unioliinae and Eragrostidinae, although *Eragrostis* is paraphyletic. Moreover, the two Zoysieae subtribes, Zoysiinae and Sporobolinae are both monophyletic.

The largest Chloridoideae tribe, Cynodonteae, is sister to Zoysieae and our sampled species here represented 19 subtribes and three genera that were previously unplaced in a subtribe. These subtribes and genera form two large sister clades: division 1 and division 2 (Figure 2-8). In division 1, Dactylocteniinae and Eleusininae form the basal-most lineage in two trees with larger numbers of genes, with Aeluropodinae occupying the next lineage; however, in three trees with smaller numbers of genes, Dactylocteniinae, Aeluropodinae, and Eleusininae form a grade outside the remaining taxa of division 1. Next, Orininae and Orcuttiinae form a grade outside a maximally supported clade containing six subtribes. In division 2, the subtribe Tripogoninae is monophyletic and sister to the other subtribes of this division. The relationships of Pappophorinae with other subtribes were different previously (Soreng et al., 2017). Among the three unplaced genera, *Kalinia* is sister to a clade of Pappophorinae, Hilariinae, and Monanthochloinae, while *Jouvea* is sister to a weakly supported clade containing Scleropogoninae + (the unplaced *Sohnsia* + Boutelouinae) in three of the coalescent trees. The well resolved relationships among *Bouteloua* species are generally consistent with previous studies (Columbus, 1999; Peterson et al., 2015).

### 2.3.5 Polyploidy in grasses and possible impact on the Poaceae phylogeny

Species with more than two sets of chromosomes are called polyploid. There are two general types of polyploid: autopolyploid with more than two (usually four) sets of chromosomes from the same species (could be from different subspecies or varieties), and allopolyploid with three or more sets of chromosomes from different species. In autopolyploid species, the chromosomes are duplicated, so are the genes, possibly followed by loss or divergence of gene copies. Whereas in allopolyploid species, genes from both genome donors are first retained and also possibly followed by gene loss (for this topic, see a review by Cheng et al., 2018). For phylogenetic studies concerning species relationship, allopolyploidy may cause a larger challenge because there might be genes from both genome donors if one were to identify orthologous genes among target species, result in some genes reflecting evolutionary history of one genome donor and other genes for the other genome donor. Grasses have experienced multiple rounds of polyploidization, including one shared by all grasses (Tang et al., 2010b), those in the early history of the Bambusoideae subfamily (allopolyploidy; Guo et al., 2019; Guo et al., 2021), and more recent ones involving members of related genera or within a genus, such as those in the tribe Andropogeneae and other grasses (Mason-Gamer et al., 2010; Liu et al., 2011; Estep et al., 2014; Triplett et al., 2014). Therefore, the bifurcating Poaceae phylogeny reported in my project is likely a simplified and limited view of the evolutionary history. Nevertheless, for the early polyploidy events shared by all grasses or many members of Bambusoideae, where the parental lineages of the presumed allopolyploid hybrids have not been identified and are possibly extinct, the relationships among grasses or bamboos, respectively, can be generally reliable, if individual orthologous group of marker genes do not include different paralogs from such allopolyploidy events. For relatively recent polyploidy events, we generally did not include more than two species affected by such an event. Thus, the effect of the recent polyploidy events on phylogeny awaits further investigation

with sampling of greater density. For Puelioideae, the proposed non-monophyly from the nuclear phylogeny could also be affected by past polyploidization, although analyses of chromosome number suggested that *Puelia* species are diploid (Dujardin, 1978; Soderstrom, 1981) and no evidence of polyploidization for *Guaduella* was reported. Furthermore, as three species in each genus were included, polyploidy in a specific member of either genus would not likely have affected the overall phylogenetic placement of these genera.

To further examine the potential effect of gene duplications (from both polyploidization events and gene-specific duplications) that occurred relatively early in the Poaceae history, we performed additional analyses. First, we obtained the local posterior probabilities (PP) and the quartet support values at each node for the three alternative topologies from coalescent analyses using the 1,150 genes and the four subsets of genes (the alternatives around a node are for a quadripartition and not the bipartition. See Supplemental Figure 4 in Huang et al., 2022) and found that most nodes have a relative high PP value and quartet support value for the first topology, including the stem nodes for monophyly of 11 subfamilies. We also found high PP support values for relationships among the subfamilies, although the quartet support was weak for some of the relationships among the PACMAD subfamilies. The generally high support for Poaceae relationships is consistent with the idea that the sub-genomes of a recent polyploid often lose genes differentially, making it more likely that the single-copy genes sampled from multiple species are derived from the same ancestral copy and thus are orthologous.

For Bambusoideae, the PP values and the quartet support values were generally supportive of the species phylogeny, including the monophyly of two tropical woody bamboo clades. It was proposed that the woody bamboos are tetraploids or hexaploids, with tropical woody bamboos and the temperate woody bamboos belonging to the tribes Bambuseae and Arundinarieae, respectively (Kellogg, 2015; Guo et al., 2019). The tropical woody bamboos form two clades, the paleotropical ones (hexaploids, including subtribes Bambusinae and Melocanninae) and the neotropical ones

(tetraploids, with the subtribes Arthrostylidiinae, Chusqueinae, and Guaduinae), while the temperate woody bamboos are also tetraploids. Guo et al. (2019) reported genomic evidence supporting a model that the paleotropical bamboos share the ABC subgenomes, the neotropical ones have the BC subgenomes, and the temperate woody bamboos carry the CD subgenomes, with diploid progenitors thought to be long extinct. The phylogeny here for Bambusoideae is consistent with the above proposed polyploid history, as the temperate, the paleotropical and the neotropical woody bamboos form separate monophyletic groups, respectively.

The quartet analyses also revealed that at some nodes there are alternative topologies with considerable support, such as those for some relationships among the PACMAD subfamilies. As these subfamilies diverged within a relatively short period of time (see below, Figure 2-9), the differences in topologies among gene trees could be due to several possible factors, such as incomplete lineage sorting and insufficient phylogenetic resolving power, in addition to the possible inclusion of paralogs from ancient polyploidy events or gene duplications. The single-gene trees were examined to reduce the effect on species phylogeny from potential paralogs and/or low-quality sequences. Firstly, I avoided genes in the larger gene sets that either have relatively low taxon coverage or short sequences and focused on the set with 436 OGs, because both low coverage and short sequences can lead to less reliable gene tree topologies. Secondly, I used monophyly of the five largest subfamilies (Bambusoideae, Chloridoideae, Oryzoideae, Panicoideae and Pooideae) as a evidence for gene orthology, as their monophyly is supported by analyses of both chloroplast genes in previous studies and the five sets of nuclear genes in my project. The examination of the gene trees for the 436 OGs suggested that among each of the 436 single gene trees, a relatively small number of sequences (1-10 sequences for 263 gene trees; 11-20 sequences for 87 gene trees; 21-43 for 51 gene trees) did not group together with the majority of sequences from the same subfamily, suggesting that they are not orthologous to most sequences of the same OG. For the 401 OGs with at least one putative non-orthologs, gene trees were reconstructed after

the removal of putative non-orthologous sequences, and a new coalescent tree was generated using the modified 436 gene set. In addition, we also generated two other coalescent trees using a 390-gene set with 90% species coverage and a 373-gene with presence in both of the two Puelioideae genera as the removal of the putative non-orthologs reduced the species coverage for some genes (see coalescent trees from filtered gene sets in Supplemental Figure 8 in Huang et al., 2022). The phylogenetic relationships in these coalescent trees are generally consistent with those in Figures 2-5, 2-6, 2-7 and 2-8; for example, the monophyly of subfamilies, the paraphyly of Puelioideae, and most of the relationships among the subfamilies are the same. Micrairoideae was supported as sister to Panicoideae in 4 of the 5 earlier coalescent trees (Supplemental Figure 3 in Huang et al., 2022), whereas it was sister to the clade of ((Chloridoideae, Danthonioideae), Arundinoideae) with support of PP=0.86 from the original 436 gene set (Supplemental Figures 4 and 8 in Huang et al., 2022). In the three coalescent trees after the removal of putative non-orthologs, Micrairoideae is sister to Panicoideae with PP values of 0.78 to 0.89 (Supplemental Figure 8 in Huang et al., 2022), suggesting that its placement next to ((Chloridoideae, Danthonioideae), Arundinoideae) was possibly influenced by non-orthologs. Therefore, although it is likely that the Poaceae history is more complex than depicted here, my results represent a major portion of the history especially at the levels of subfamily and tribe.

### 2.3.6 Lower cretaceous origin of Poaceae

The newly reconstructed nuclear phylogeny of Poaceae provides an opportunity to estimate the origin and divergence times of major lineages using molecular clock analysis. Early studies suggested that there are a few pollen fossils related to Poaceae dated to no older than 70 my (million years; Muller, 1981). More recently, earlier fossils have been discovered (Shi et al., 2012; Wu et al., 2018), supporting older ages for Poaceae and its major clades, with an estimate of Poaceae

crown age to be older than 100 my using these fossil calibrations (Schubert et al., 2019). I performed molecular clock analysis using TreePL 1.0 (Smith and O'Meara, 2012) with the ML tree from concatenated super-matrix of 180 nuclear genes. To include fossil calibrations outside the Poaceae family, six more outgroup species were included in addition to the seven species in the coalescent analyses. The upper or/and lower boundaries of these calibrations were set according to previous studies (see Appendix A).

Results show the crown age of Poaceae to be ~101 my, in the Lower Cretaceous (Figure 2-9 shows mean values of ages; see Appendix B for confidence intervals). Following the successive divergences of the four early lineages of Poaceae spanning a period of ~20 my, the crown age of (PACMAD + BOP) is estimated to be ~81 my in the Upper Cretaceous. Thus, the PACMAD and BOP clades probably diverged before the Cretaceous-Paleogene (K-Pg) boundary. Furthermore, the three BOP subfamilies also diverged from each other before the K-Pg boundary (between ~78-74 mya), whereas the six PACMAD subfamilies separated from each other shortly after the K-Pg boundary during a period of less than 7 my (~66.33-59.86 mya). Subsequently, most tribes in large subfamilies diverged over much longer periods. For example, among the tribes in Pooideae, Brachyelytreae (*Brachyelytrum*) diverged from other Pooideae at ~67 mya, whereas Bromeae (*Bromus*) separated from Triticeae (with *Leymus* and *Triticum*) much more recently at ~19 mya. Thus, the divergence among tribes of Pooideae spanned a period of ~48 my. Similarly, the tribes in Panicoideae diverged over a period of ~30 my (from ~54.22 to ~23.54 mya). Our analyses also provide divergence times for subtribes and genera, showing a general tendency for the divergence times in Oryzoideae and Bambusoideae to be older than those in Pooideae, Panicoideae, and Chloridoideae.

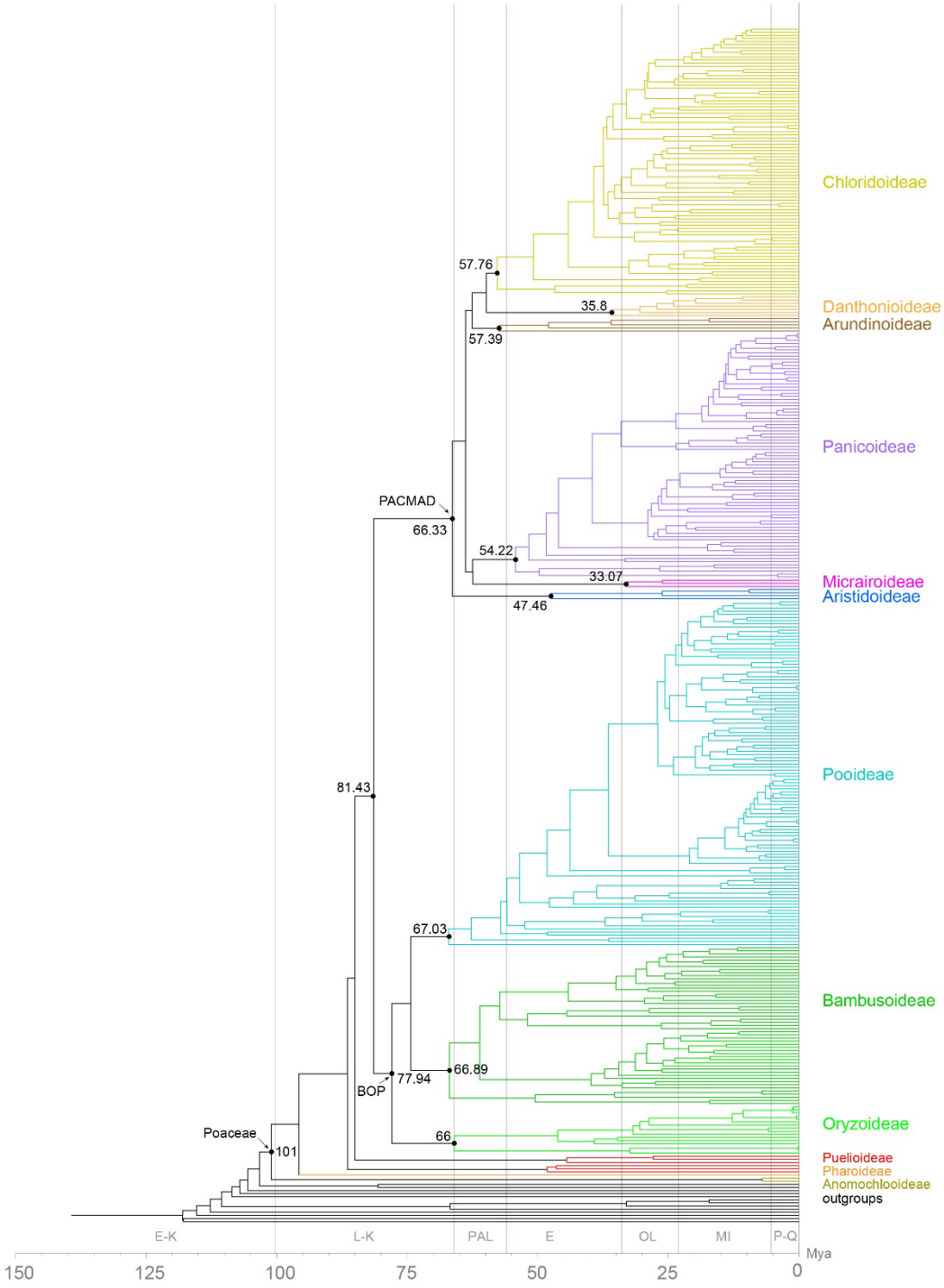


Figure 2-9: Divergence time estimation for Poaceae. A Poaceae time tree was estimated by treePL using the topology of the 1150-gene coalescent tree as a backbone and the concatenated super-matrix from 180 genes for calculation of branch length. Subfamily names are lablled to the right with different colors. Geological time scale is shown at the bottom, and periods are delimited with vertical gray lines. E-K, early Cretaceous; L-K, late Cretaceous; PAL, Paleocene; E, Eocene; OL, Oligocene; MI, Miocene; P-Q, Pliocene and Quaternary. Estimated divergence times of subfamilies are marked at corresponding nodes, and confidence intervals are listed in Appendix B.

### 2.3.7 Ancestral character reconstruction supports multiple origins of $C_4$ photosynthesis in PACMAD grasses

The success of grasses is thought to be due in part to their ability to fix carbon via $C_4$ photosynthesis, which facilitates adaptation to habitats with stressful conditions, such as high temperature or light intensity, aridness, and salinity (Christin et al., 2007a; Edwards and Still, 2008). Under hot and dry environments, plants tend to close stomata to retain water, leading to reduced access to $CO_2$; many $C_4$ plants have evolved a specialized organization of leaf tissues called Kranz anatomy (Tregunna et al., 1970; Smith and Epstein, 1971) that could increase local concentration of $CO_2$ near the carbon-fixing enzyme Rubisco, by physically separating the light-dependent reactions and the Calvin cycle.

Among families with $C_4$ photosynthesis, Poaceae has the largest number (~4,500 species, ~60% of all $C_4$ plants) of $C_4$ species; $C_4$ species are also reported in several other angiosperm families: Amaranthaceae, Asteraceae, Brassicaceae, Cyperaceae, and Euphorbiaceae. All $C_4$ Poaceae belong to the PACMAD clade, although they do not form a monophyletic group. To date, four subfamilies are reported to have $C_4$ species, namely Aristidoideae, Micrairoideae, Panicoideae and Chloridoideae. However, the existence of undiscovered $C_4$ species in Arundinoideae or Danthonioideae cannot be excluded, since the photosynthetic pathway is somewhat a continuous, complex trait and sometimes there are both $C_3$ and $C_4$ ecotypes/subspecies within a species, such as *Alloteropsis semialata* (Lundgren et al., 2016).

Among Poaceae species sampled for this project, 150 have been described as $C_4$ species, including one in Micrairoideae, three in Aristidoideae, 62 in Panicoideae, and 84 in Chloridoideae, according to Soreng et al. (2017). The $C_3$/$C_4$ photosynthetic ancestral states were reconstructed using a Most Parsimony method implemented in Mesquite (version 3.6), with the coalescent trees from five different gene sets. As $C_4$ species are only known in the four PACMAD subfamilies, our analyses support the hypotheses that the most recent common ancestor (MRCA) of Poaceae, the

four nodes for the separation of the three earliest divergent subfamilies, the crown node of BOP + PACMAD, and the MRCAs of the BOP clade as well as the three BOP subfamilies were all $C_3$.

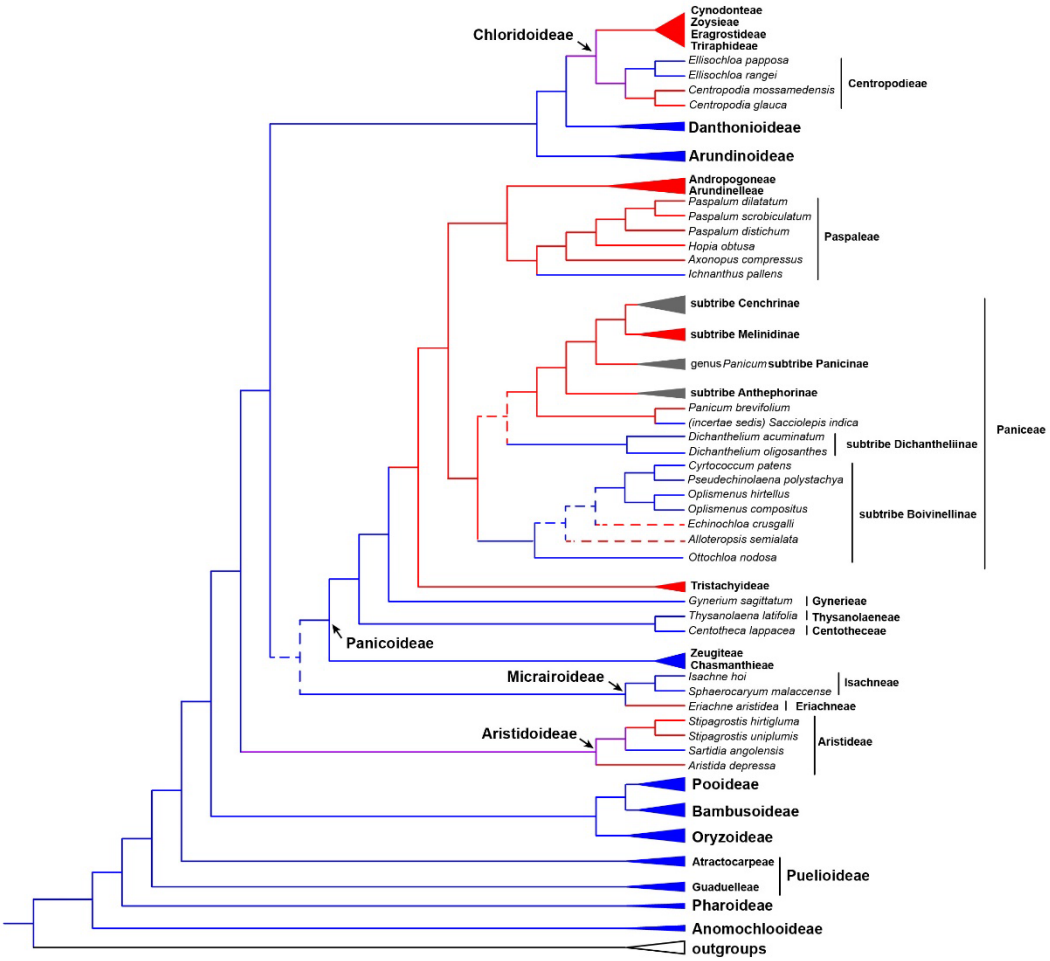


Figure 2-10: Ancestral state reconstruction of photosynthetic pathway type in Poaceae. Ancestral states of photosynthetic pathway type ($C_3$/$C_4$) were estimated using information from extant taxa (species) by the maximum-likelihood method using the Mesquite program. Terminals and branches are marked with different colors to represent photosynthetic type. Red indicates $C_4$, blue indicates $C_3$, purple indicates an uncertain ancestral state, and gray indicates a mixed clade that includes both $C_3$ and $C_4$ species/genera, although some were not sampled. The mixed clades shown here are dominated by $C_4$ species, and our sampling included only the $C_4$ species. Branches associated with alternative topology are shown in dashed lines.

The subfamily Aristidoideae is sister to the other PACMAD subfamilies, consistent with the relationship summarized recently (Soreng et al., 2017). The MRCA of PACMAD and that of the five subfamilies after the divergence of Aristidoideae are both proposed to be $C_3$ (Figure 2-10). However, the ancestral state of Aristidoideae is uncertain. All three genera in Aristidoideae are sampled, with a phylogenetic topology of *Aristida* being sister to (*Sartidia* + *Stipagrostis*), consistent with previous studies (Cerros-Tlatilpa and Columbus, 2009). The fact that most *Aristida* and *Stipagrostis* species are $C_4$ while *Sartidia* and *Aristida longifolia* are $C_3$ makes the ancestral state of Aristidoideae equivocal. If the Aristidoideae ancestor was $C_3$, then $C_4$ has originated at least twice, in *Aristida* and *Stipagrostis*, respectively, while *Sartidia* has retained the ancestral state, consistent with a previous report (Cerros-Tlatilpa and Columbus, 2009). A less likely scenario would be that the Aristidoideae ancestor was already $C_4$, and there were reversions to $C_3$ in *Sartidia* and *Aristida longifolia*.

The ancestral state of each of the other three subfamilies with $C_4$ species is proposed to be $C_3$, according to the reconstruction here. More specifically, Micrairoideae was estimated to be $C_3$ in all analyses from five different coalescent trees, regardless of the placement of this subfamily. Our sampling includes three Micrairoideae genera and the $C_4$ genus *Eriachne* is sister to the clade of *Isachne* + *Sphaerocaryum* (both are $C_3$); alternatively, the ancestral state of Micrairoideae could also be $C_4$, with a reversion to $C_3$ in the MRCA of *Isachne* + *Sphaerocaryum*. Among the genera in Micrairoideae not sampled here, *Micraira* alone defines a tribe (Micraireae) and was placed as the first divergent lineage in the subfamily (Soreng et al., 2017). As *Micraira* is $C_3$, its placement as sister to the other Micrairoideae genera would support $C_3$ as the ancestral state of Micrairoideae.

Panicoideae is the largest subfamily in PACMAD with 13 tribes and ten tribes are represented by our sampling, including five with $C_3$ species (Centotheceae, Chasmanthieae, Gynerieae, Thysanolaeneae, and Zeugiteae), three with only $C_4$ species (Andropogoneae, Arundinelleae, and Tristachyideae), and two with both $C_3$ and $C_4$ species (Paniceae and Paspaleae).

The three small tribes not sampled here [Cyperochloeae (two species), Lecomtelleae (one species), and Steyermarkochloeae (two species)] have only $C_3$ species. In the phylogeny here, the earliest divergent lineage of Panicoideae has two tribes, Chasmanthieae and Zeugiteae; the sister relationship of Chasmanthieae and Zeugiteae is consistent with previous reports (Soreng et al., 2017; Saarela et al., 2018). The next two early separating branches are (Centotheceae + Thysanolaeneae) and Gynerieae, respectively. The above mentioned three early divergent lineages are all $C_3$; thus, this relationship strongly supports $C_3$ as the ancestral state of Panicoideae. The MRCA of the other five tribes sampled here, and the MRCA of each of these tribes (Tristachyideae, Paniceae, Paspaleae, Arundinelleae, and Andropogoneae) are supported by the ancestral character reconstruction to be $C_4$, even though Paniceae and Paspaleae also have $C_3$ species (Figure 2-10). The three small $C_3$ tribes not sampled here were previously placed outside this large clade of five tribes (Soreng et al., 2017); therefore, their placements also support $C_3$ as ancestral for Panicoideae. Within Paniceae, the inferred ancestral state of subtribes Boivinellinae and Dichantheliinae varies between analyses, mainly due to uncertainties in the phylogeny (Supplemental Figures 3, 4 & 10 in Huang et al., 2022). Nevertheless, our analyses support two likely $C_4$ to $C_3$ reversions for the two $C_3$ species *Ichnanthus pallens* and *Sacciolepis indica*, in tribes Paspaleae and Paniceae, respectively. The mostly $C_4$ Paspaleae is sister to a combined clade of two $C_4$ tribes, Andropogoneae and Arundinelieae, supporting the ancestor of Paspaleae to be $C_4$ and a reversion to $C_3$ in *Ichnanthus pallens.* In Paniceae, *S. indica* was previously not assigned to a subtribe, but in our results, it is supported as sister to the $C_4$ species *Panicum brevifolium*; this relationship supports a reversion to $C_3$ in *S. indica*. Three subtribes in Paniceae: Anthephorinae, Cenchrinae, and Panicinae, contain both $C_3$ and $C_4$ species, but our sampling was incomplete; in addition, the tribe Paspaleae contains a few other $C_3$ genera that are not included here. Therefore, more sampling is needed to better understand the evolution of $C_4$ photosynthesis in these two tribes.

Among the remaining three subfamilies of the PACMAD clade, Arundinoideae and Danthonioideae are entirely $C_3$ and are placed as successive sisters of Chloridoideae, supporting the MRCA of the combined clade of these three subfamilies as $C_3$. Within Chloridoideae, the tribe Centropodieae has two genera, *Centropodia* ($C_4$) and *Ellisochloa* ($C_3$), and is sister to all the other chloridoid grasses, making the ancestral states of Centropodieae and Chloridoideae uncertain in our analysis, even though the MRCA of the combined clade of the other four tribes is inferred to be $C_4$. If the MRCA of Chloridoideae was $C_3$, then there were two origins of $C_4$ photosynthesis, one for *Centropodia* and the other for the MRCA of the other four tribes with the majority of Chloridoideae. However, if the ancestral states of Chloridoideae and Centropodieae were both $C_4$, then there was one origin of $C_4$ for the subfamily and a reversion to $C_3$ for *Ellisochloa.*

The ancestral state reconstruction analyses support independent origins of $C_4$ in each of the four subfamilies with $C_4$ species, possibly twice in Aristidoideae and Chloridoideae, respectively, and multiple times in Panicoideae. This is consistent with a previous report by GPWG II. (2012), which proposed as many as 24 separate transitions from $C_3$ to $C_4$ photosynthesis, among which two were in Aristidoideae, two in Chloridoideae, one in Micrairoideae and the rest all in Panicoideae. Notably, there were possible reversions from $C_4$ back to $C_3$ in Panicoideae (tribe Paniceae).

It should be noted that sampling limitations here, including the lack of some $C_3$ lineages, probably affected some of the ancestral state reconstruction results. In Panicoideae, especially in supertribe Panicodae and Andropogonodae, the sampling favored $C_4$ species, likely increasing the probability of inference of the ancestral nodes of these two supertribes as $C_4$. On the other hand, our sampling included early diverging tribes of Panicoideae, such as Chasmanthieae and Zeugiteae, which are $C_3$, supporting the inferred ancestral state of Panicoideae as $C_3$. Previous studies on Panicoideae phylogeny mostly using plastome genes reconstructed different relationships among some tribes, especially for basal ones (e.g., position of Tristachyideae, $C_4$). Therefore, although our sampling is indeed incomplete at subtribe level, our well-supported phylogeny provides meaningful

information. Additional studies with greater sampling are needed to investigate the previously proposed >20 transitions from $C_3$ to $C_4$ (GPWG II., 2012) and to resolve relationships among some Paniceae subtribes. Moreover, $C_4$ photosynthesis is a complex trait with changes in both leaf anatomy and biochemical processes (GPWG II., 2012; Washburn et al., 2015); thus, even closely related $C_4$ species might have experienced distinct evolutionary histories for $C_4$ photosynthesis, as noted previously (Sinha and Kellogg, 1996; Christin et al., 2010; Dunning et al., 2017; Moreno-Villena et al., 2018) and also supported by the evolutionary analyses of homologs of *ppc* genes in the next section.

### 2.3.8 Phylogenetic analyses of the *ppc* gene family provide molecular evidence for independent origins of $C_4$ photosynthesis in grasses

The $C_4$ photosynthesis processes depend on multiple genes that are responsible for biochemical pathways and leaf anatomy and are co-opted for the $C_4$ functionality (Moreno-Villena et al., 2018). Among these genes, the *ppc* gene that encodes PEPC (phospho*enol*pyruvate carboxylase), which is responsible for the initial fixation of atmospheric $CO_2$ into organic compounds (Sage, 2004), has been studied in several plant families including Poaceae, Asteraceae and Fabaceae (Bläsing et al., 2000; Christin et al., 2007a; Christin and Besnard, 2009; Wang et al., 2016). The *ppc* gene belongs to a gene family encoding several enzymes involved in photosynthesis and some stress-response processes. Previous studies of the *ppc* family indicated that *ppc* genes for $C_4$ photosynthesis encode proteins with shared sequence motifs (Bläsing et al., 2000; Christin et al., 2007a; Paulus et al., 2013) and that the $C_4$ *ppc* genes in Poaceae originated from non-$C_4$ paralogs in two different *ppc* clades (Christin and Besnard, 2009), sometimes involving possible horizontal gene transfer (Christin et al., 2012). Previous phylogenetic analysis of *ppc* gene sequences from several Poaceae species and other Poales (*Eleocharis*, Cyperaceae), other monocots (*Aloe*,

Asphodelaceae; *Hydrilla*, Hydrocharitaceae; *Vanilla*, Orchidaceae) and several eudicot families helped to define several gene clades for grass *ppc* genes (here named as subclades): *ppc-aL1a, ppc-aL1b,ppc-aL2, ppc-B1, ppc-B2*, and *ppc-aR* (Christin and Besnard, 2009). However, the origins of these subclades were not clear.

To gain additional insights into the evolution of $C_4$ photosynthesis in the PACMAD clade, phylogenetic analyses of the Poaceae *ppc* gene family were performed with putative *ppc* genes from all grass subfamilies, except Puelioideae (with only low-coverage genome sequences), and nine of 15 non-grass Poales families that represent all major Poales clades. In addition, representatives of Musaceae (Zingiberales) and Asparagaceae (Asparagales) were included as outgroups. Our phylogenetic results support a model that the grass *ppc* subclades originated first with a duplication shared by the MRCA of both Poales and Zingiberales (Figure 2-11, indicated by one of the stars), after divergence from Asparagales. Subsequently, another duplication in the early Poales history, probably after the separation of Typhaceae, generated the common ancestor of both the *ppc-aL1a* and *ppc-aL1b* subclades, and the ancestor of the *ppc-aL2* subclade, while a later duplication likely at the MRCA of Poaceae produced the *ppc-aL1a* and *ppc-aL1b* subclades (Figure 2-11A). Although the origins of the *ppc-B1*, *ppc-B2*, and *ppc-aR* subclades are less clear, they seem to result from duplications of an ancestral gene shared by Poales and Zingiberales (Musaceae), after the divergence of most families in Poales. However, the placements of a *ppc-B1*-like gene from Flagellariaceae, a family closely related to Poaceae, and genes from Anomochlooideae, the grass subfamily that is sister to all other Poaceae, indicate that further analysis is needed with genes from more representatives of families closely related to Poaceae to resolve the early histories of the *ppc-B1*, *ppc-B2*, and *ppc-aR* subclades.

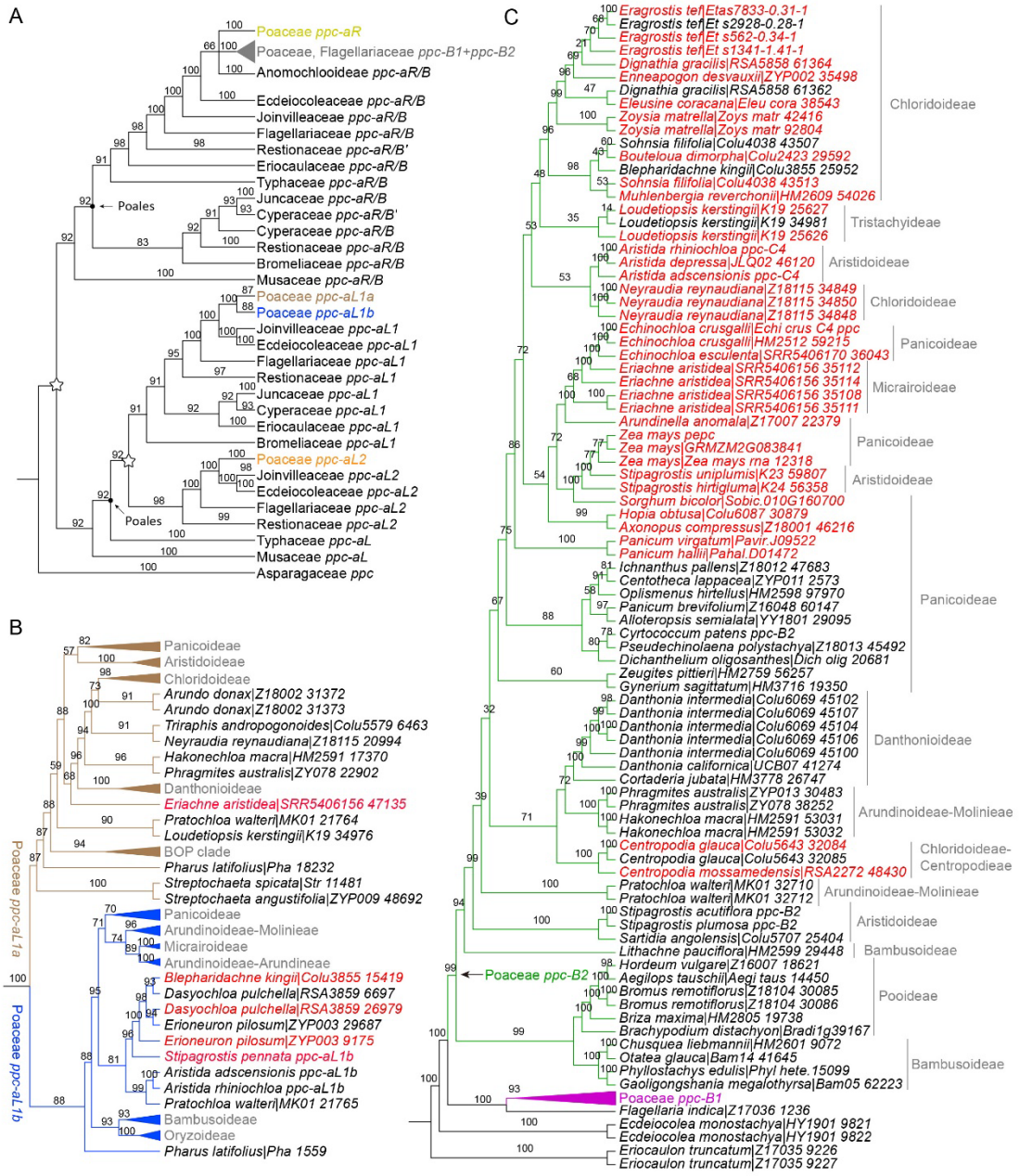


Figure 2-11: Molecular phylogenetic analyses of the *ppc* gene family. (A) A simplified presentation of the *ppc* gene family from a maximum-likelihood analysis using 516 nucleotide sequences of protein coding regions. (B) An illustration of subclades *ppc-aL1a* and *ppc-aL1b* in the *ppc* tree shown in (A). Names of sequences with a putative $C_4$ function are marked in red, and non-$C_4$ sequences are marked in black. Bootstrap values are shown above branches. (C) Summary of amaximum-likelihood gene tree based on 119 nucleotide coding sequences from Poaceae *ppc-B1* and *ppc-B2* only, plus four outgroup sequences. Subfamily names are marked to the right of gene clades, as is the tribe name Tristachyideae (Panicoideae).

Previous comparative analyses of PEPC amino acid sequences from Poaceae and other families for $C_4$ photosynthesis or non-$C_4$ functions revealed characteristic residues at multiple positions (Christin et al., 2007a; Christin and Besnard, 2009). The available sequences from many Poaceae members provide an opportunity to further examine the conservation of these residues; our comparison of over 500 PEPC sequences indicated that characteristic residues for either putative $C_4$ or non-$C_4$ enzymes are very similar to those reported previously (Supplemental Table 9 in Huang et al., 2022). Also, it was reported that in some $C_4$ plants the *ppc* gene for $C_4$ photosynthesis was expressed at higher levels (Moreno-Villena et al., 2018). To investigate whether the putative $C_4$ *ppc* genes identified here were also more highly expressed, we examined expression level by mapping RNA-seq reads to mRNA sequences of different *ppc* genes. Our results suggested that, for some species, the putative $C_4$ *ppc* gene was likely expressed at a higher level than other *ppc* genes in the same species, such as those in *Centropodia glauca*, *Neyraudia reynaudiana, Eriachne aristidea, Loudetiopsis kerstingii, Echinochloa esculenta*, and *Hopia obtusa*; however, in several other species, the putative $C_4$ *ppc* genes appeared not to be the most highly expressed ones (Supplemental Table 10 in Huang et al., 2022). It is possible that the transcriptomes of different species contain different amounts of photosynthetic organs/tissues and more detailed information about *ppc* gene expression is needed to understand expression patterns of $C_4$ and non-$C_4$ *ppc* genes.

The *ppc* gene phylogenetic analysis here also indicates that the putative *ppc* genes with the characteristic motif for $C_4$ photosynthesis belong to one of three subclades: *ppc-aL1a*, *ppc-aL1b*, and *ppc-B2*, strongly supporting the hypothesis of multiple $C_4$ origins in Poaceae. Specifically, among the three genera in Aristidoideae, both *Aristida* and *Stipagrostis* have numerous $C_4$ species, whereas *Sartidia* has only $C_3$ species. Here *ppc* genes were identified from members of each of the three genera, including non-$C_4$ *ppc* genes from *Aristida* (*A. adscensionis* and *A. rhiniochloa*), *Sartidia angolensis, Stipagrostis* (*S. acutiflora* and *S. plumosa*) and $C_4$ *ppc* genes from *Aristida* and *Stipagrostis* species. Previously it was reported that the $C_4$ genes in *Aristida* (*A. adscensionis* and

*A. rhiniochloa*) were in the *ppc-B2* subclade, whereas the $C_4$ genes in *Stipagrostis pennata* were in *ppc-aL1a* (Christin and Besnard, 2009), indicating that *ppc* genes for $C_4$ photosynthesis probably originated at least twice in Aristidoideae. Our results suggest that the evolution of $C_4$ *ppc* genes in Aristidoideae might be more complex; in addition to confirming the previous findings, our analyses identified $C_4$ *ppc* genes from two *Stipagrostis* species not sampled previously (*S. hirtigluma* and *S. uniplumis*) and a third *Aristida* species (*A. depressa*) as belonging to the *ppc-B2* subclade.

In Micraioideae, *Eriachne* is the only $C_4$ lineage, with five identified *ppc* sequences predicted to be functionaly $C_4$; one of these is placed in *ppc-aL1a* as being related to $C_3$ *ppc* genes from several other PACMAD subfamilies, suggesting an independent origin of $C_4$ *ppc*, which was not previously reported. Four other *ppc* sequences from *Eriachne* are placed in *ppc-B2*, and they are all closely related to sequences from two *Echinochloa* species. This relationship is further supported by a BLAST search showing that the most similar sequences of these four *Eriachne ppc-B2* sequences are from *Echinochloa*, which is a $C_4$ species in the tribe Paniceae of Panicoideae. Furthermore, the $C_4$ *ppc* genes from both *Echinochloa* and *Eriachne* are close to $C_4$ genes from other members of Panicoideae. One possible explanation for this observation is horizontal gene transfer between *Echinochloa* and *Eriachne*, although convergent evolution cannot be ruled out. Although not yet available, genomic sequences of these $C_4$ *ppc* genes, for example, may provide information regarding whether there was horizontal gene transfer between *Echinochloa* and *Eriachne*, as intronic and untranslated regions are less likely to evolve convergently. On the other hand, if introns are missing in the genomic sequences, that would be a strong evidence of horizontal gene transfer mediated by RNA (Kim et al., 2014).

Panicoideae and Chloridoideae are the two largest PACMAD subfamilies and contain the majority of $C_4$ species in Poaceae, although only a subset was included in the *ppc* gene family analysis here. In Panicoideae, all $C_4$ *ppc* genes identified here (from *Arundinella*, *Axonopus*, *Digitaria*, *Hopia*, *Loudetiopsis*, *Zea mays*, and *Echinochloa*) are in the *ppc-B2* subclade. As

mentioned in the previous section, our ancestral character analyses identified two possible $C_4$ to $C_3$ reversions (or retentions of ancestral state $C_3$) in the Panicoideae members *Ichnanthus pallens* and *Sacciolepis indica*. No $C_4$ type *ppc* sequences were found from the transcriptomes of these two species, providing further support for their $C_3$ state. In Chloridoideae, the *ppc* gene family analysis showed that the $C_4$ *ppc* genes in several Chloridoideae genera (*Bouteloua*, *Centropodia*, *Dignathia*, *Enneapogon*, *Neyraudia*, *Muhlenbergia*, *Sohnsia*, and *Zoysia*) are in the *ppc-B2* subclade. On the other hand, $C_4$ *ppc* genes in three closely related genera (*Blepharidachne, Dasyochloa*, and *Erioneuron,* all in the subtribe Scleropogoninae of the tribe Cynodonteae; Figure 2-9) are in the *ppc-aL1b* subclade, supporting a different origin. The *ppc* gene phylogeny here is different from the species phylogeny, probably caused by convergent evolution of $C_4$ *ppc* genes included here, as proposed by Christin et al (2007a). In addition, horizontal gene transfer was also hypothesized for *ppc* in *Alloteropsis* species in Panicoideae (Christin et al., 2012), and is a possible explanation for the cases of *Eriachne aristidea* and *Echinochloa* $C_4$ *ppc* genes.

## 2.4 Discussion

### 2.4.1 A well-resolved Poaceae nuclear phylogeny supporting monophyly of most subfamilies and tribes

I present a generally well-resolved Poaceae phylogeny that supports the monophyly of 11 out of 12 subfamilies and most of the tribes with two or more sampled taxa, largely consistent with recent classifications (Kellogg, 2015; Soreng et al., 2015; Soreng et al., 2017). In addition, the nuclear phylogeny provides well-resolved relationship among subfamilies and also for many tribes and some subtribes. Specifically, the deep relationships in the PACMAD clade have long been difficult to resolve and various topologies have been reported using chloroplast and mitochondrial genes or a small number of nuclear genes (GPWG II, 2012; Christin et al., 2014; Soreng et al., 2015; Soreng et al., 2017; Saarela et al., 2018), with conflicts between results from chloroplast and mitochondrial genes (Cotton et al., 2015) or among different sets of genes (Saarela et al., 2018). Although previous studies placed (Chloridoideae + Danthonioideae) as sister to (Arundinoideae + Micrairoideae) (Soreng et al., 2017; Saarela et al., 2018), both Aristidoideae and Panicoideae was reported to be the first divergent lineage among the PACMAD subfamilies. The branches subtending the individual PACMAD subfamilies are usually short, suggesting rapid diversification among these subfamilies. Our analyses from both coalescence and super-matrix estimated Aristidoideae as sister to the remaining PACMAD subfamilies, and Arundinoideae as sister to (Danthonioideae + Chloridoideae). Micrairoideae is supported in most coalescent analyses as sister to Panicoideae, whereas the signal for its placement as sister to (Arundinoideae + (Danthonioideae + Chloridoideae)) might be due to paralogous sequences possibly generated by ancient polyploidization events.

### 2.4.2 Phylogenetic analysis of *ppc* gene family provides insights into evolution of $C_4$ photosynthesis

Six subclades for grass *ppc* genes: *ppc-aL1a*, *ppc-aL1b*, *ppc-aL2*, *ppc-B1, ppc-B2,* and *ppc-aR,* were defined by previous molecular phylogenetic analysis using sequences from several Poaceae species, other Poales (*Eleocharis*, Cyperaceae), other monocots (*Aloe*, Asphodelaceae; *Hydrilla*, Hydrocharitaceae; *Vanilla*, Orchidaceae) and eudicots (Christin et al., 2007a; Christin and Besnard, 2009). However, whether these subclades were produced by polyploidy events or specific gene duplications was not clear. Also, other distant *ppc* paralogs exist but are not closely related to genes known for function in photosynthesis (Moreno-Villena et al., 2018) and not analyzed here. The analysis here included a broad sampling of *ppc* homologs from Poaceae and 10 other Poales families, as well as Musaceae (Zingiberales) and Asparagaceae (Asparagales), providing better understanding of the early histories of the grass *ppc* genes. The results indicate that the six grass *ppc* subclades belong to two ancient clades, *ppc-aL* and *ppc-aR/B* (each with three subclades), and both likely originated in the MRCA of Poales and Zingiberales. The duplication of the ancestral *ppc-aL* gene in early Poales generated the *ppc-aL1* and *ppc-aL2* clades and a subsequent duplication of *ppc-aL1* in early Poaceae produced the *ppc-aL1a* and *ppc-aL1b* subclades. However, the evolution of *ppc-aR/B* genes to *ppc-aR*, *ppc-B1* and *ppc-B2* subclades is less clear, although one possible scenario is that a duplication in the MRCA of Poaceae generated the *ppc-aR* and *ppc-*(*B1*+*B2*) clades.

The putative $C_4$ *ppc* genes were identified by the conserved Serine residue (corresponding to residue #780 in the *Zea mays* PEPC, GRMZM2G083841) and mostly belong to the *ppc-*(*B1*+*B2*) clade, with a few in the *ppc-aL1a* and *ppc-aL1b* subclades. Previously, the *ppc-*(*B1*+*B2*) genes formed two subclades (Christin et al., 2007a; Christin and Besnard, 2009). Here a phylogenetic analysis of the *ppc-*(*B1*+*B2*) genes also yielded two highly supported clades, *ppc-B1* and *ppc-B2* (Figure 2-11 C), with known *ppc-B1* and *ppc-B2* genes, respectively (Christin et al., 2007a; Christin

and Besnard, 2009). To avoid possible effects of natural selection on gene phylogeny, another analysis using the nucleotide residues at the third codon positions was also performed; although detailed phylogenetic relationships among gene sequences are somewhat different, both *ppc-B1* and *ppc-B2* clades were recovered and $C_4$ sequences were clustered in the *ppc-B2* clade (Supplemental Figure 14 in Huang et al., 2022). The *ppc-B1* clade contains genes from the early-divergent Poaceae subfamilies Anomochlooideae and Pharoideae and both the BOP and PACMAD clades. The *ppc-B2* subclade includes most of putative $C_4$ *ppc* genes, as well as non-$C_4$ *ppc-B2* homologs from several BOP and PACMAD subfamilies, but not Oryzoideae and the early-divergent subfamilies. Therefore, *ppc-B1* and *ppc-B2* subclades likely resulted from a duplication in the MRCA of Poaceae, but *ppc-B2* genes were lost from (or not expressed in) members of several subfamilies sampled here, all containing $C_3$ plants.

In the *ppc-B2* clade, most putative $C_4$ genes are clustered into one large clade, except for $C_4$ genes from the early divergent Centropodieae of Chloridoideae, suggesting that the Centropodieae $C_4$ genes had a separate origin from the other $C_4$ genes in the *ppc-B2* clade; a putative origin of $C_4$ photosynthesis in *Centropodia* distinct from other Chloridoideae is also supported by ancestral character reconstruction (Figure 2-10). Most of the other $C_4$ *ppc-B2* genes form a clade with 86% BS support, suggesting that they might have a single origin; however, their relationships do not agree with the species relationships, as noted previously (Christin et al., 2007a). The relationships among the subfamilies in the PACMAD clade were difficult to resolve even using multiple genes (Prasad et al., 2011; Soreng et al., 2017). Therefore, it is not surprising that the $C_4$ *ppc-B2* genes do not follow the species relationships. Nevertheless, $C_4$ *ppc-B2* genes of two Aristidoideae genera (*Aristida* and *Stipagrostis*) were placed, respectively, close to genes from Panicoideae (maize and sorghum) and Chloridoideae species (in Cynodonteae and three other tribes), suggesting that the *Aristida* and *Stipagrostis* $C_4$ genes might not be from the same origin in the subfamily.

Previously, a *Stipagrostis pennata* $C_4$ gene in the *ppc-aL1b* subclade (Christin and Besnard, 2009) supported a different origin from that for $C_4$ *ppc-B2* genes in *Stipagrostis hirtigluma* and *S. uniplumis.* We also identified $C_4$ *ppc-aL1b* genes from three other species (*Blepharidachne kingii*, *Dasyochloa pulchella*, and *Erioneuron pilosum*) belonging to the same subtribe, Scleropogoninae, in the tribe Cynodonteae of Chloridoideae (Figure 2-11 B), suggesting a shared $C_4$ origin in the ancestor of the subtribe Scleropogoninae. The close relationship of these genes to that from *Stipagrostis pennata* in Aristidoideae might be explained by horizontal transfer between respective members of Aristidoideae and Cyndonteae (Chloridoideae). Moreover, a putative $C_4$ gene was identified in the *ppc-aL1a* subclade from *Eriachne aristidea* of Micrairoideae (Figure 2-11 B); this species also has $C_4$ *ppc-B2* gene(s) related to those from Panicoideae species, suggesting that the $C_4$ *ppc-B2* genes might have experienced horizontal transfer from a Panicoideae taxa to *Eriachne aristidea*, as proposed previously among Paniceae members (Christin et al., 2012).

The phylogenetic analyses of *ppc* homologs expanded the coverage of subfamilies compared to previous studies to include all subfamilies, except Puelioideae, and identified more putative grass $C_4$ genes belonging to the *ppc-B2* and *ppc-aL1b* subclades; in addition, the analyses here uncovered a new $C_4$ gene of the *ppc-aL1a* subclade. The results support at least three origins of $C_4$ genes in Chloridoideae (two in *ppc-B2* and one in *ppc-aL1b*), at least three origins in Aristidoideae (two in *ppc-B2* and one in *ppc-aL1b*), at least two origins in Micrairoideae (one in *ppc-B2* and one in *ppc-aL1a*) and multiple times in Panicoideae. These findings indicate that there were not only multiple origins of $C_4$ *ppc*, but that members of at least three *ppc* subclades were recruited. The clades containing most $C_4$ species in both Panicoideae (with large tribes Andropogoneae and Paniceae) and Chloridoideae (with the largest tribe Cynodonteae) originated during the early to middle Eocene (Figure 2-9). As the Earth's temperature was relatively high during this period, the evolution of $C_4$ photosynthesis might have promoted adaptation to warm

environments and contributed to the diversification of Andropogoneae/Paniceae and Cynodonteae in the two largest PACMAD subfamilies.

# Chapter 3

# An expanded Panicoideae phylogeny and evolution of $C_4$ related gene *ppc*

Panicoideae is a large and diverse subfamily, with economically important crop species. The complexity of photosynthetic pathway evolution is especially exemplified in this subfamily, with a mixture of $C_3$ and $C_4$ species in a couple of tribes. In this chapter, Panicoideae phylogeny is resolved using low-copy nuclear data from both transcriptomic and genomic datasets, and an expanded gene family analysis of *ppc* is performed to further investigate how different paralogs contribute to the origin of $C_4$ pathway.

## 3.1 Introduction and objectives

In this section, general information about the subfamily Panicoideae will be introduced. Phylogeny of Panicoideae from previous studies is summarized and compared, and this includes relationship among tribes and subtribes. Objectives for this part of the project relevant to Panicoideae phylogeny are stated.

### 3.1.1 Current phylogeny of Panicoideae

Panicoideae is one of the six subfamilies in the PACMAD clade of Poaceae, with 14 tribes and over 3,300 species, including important crops such as maize (*Zea mays*), sorghum (*Sorghum* spp), sugarcane (*Saccharum officinarum*) and foxtail millet (*Setaria italica*). Earlier taxonomic studies recognized the three tribes Andropogoneae, Paspaleae and Paniceae as Panicoideae *s.s.* (strict sense, or *sensu stricto*). The following seven tribes, Chasmanthieae, Zeugiteae,

Steyermarkochloeae, Tristachyideae, Centotheceae, Cyperochloeae and Thysanolaeneae were formerly organized into a subfamily “Centothecoideae” or put into other Panicoideae tribes based on morphological data. Tribe Gynerieae with only one genus, *Gynerium*, was first proposed to be as part of Panicoideae rather than Arundinoideae by Sánchez-Ken and Clark (2001). Tribe Arundinelleae, consists of *Arundinella* and *Garnotia*, is treated as a subtribe of Andropogoneae in some literature (Kellogg, 2015) because of its small size and being sister to Andropogoneae; but most others recognized it as a separate tribe (Aliscioni et al., 2012; Soreng et al., 2017; Saarela et al., 2018). Similarly, the genus *Lecomtella* was not recognized as a tribe by Kellogg (2015), although decades ago the name Lecomtelleae was already proposed by Pilg ex. Potztal (1957). Both Soreng et al. (2017) and Saarela et al. (2018) reported *Lecomtella* to be a separate lineage sister to Panicoideae s.s. Tribe Jansenelleae (*Jansenella* and *Chandrasekharania*) was recently proposed by Bianconi et al. (2020) using both plastid and nuclear phylogenies. These two genera, although sister to Andropogoneae which is completely $C_4$, are confirmed to be $C_3$ by leaf anatomy and carbon isotope analysis.

Possibly due to different sampling and methodology, the relationship among Panicoideae tribes has been incongruent among studies. As exemplified in Figure 3-1, most studies supported the same topology within Panicoideae *s.s.*, where Paniceae is sister to (Paspaleae, (Andropogoneae, Arundinelleae)). As for *Lecomtella*, Besnard et al. (2013) identified it as an isolated lineage in Panicoideae but failed to resolve its exact position using plastid and nuclear genes. Soreng et al. (2017) and Saarela et al. (2018) placed it sister to Panicoideae s.s. based on plastid genes with more extensive sampling of taxa. Nevertheless, the position of *Lecomtella* awaits to be verified by a larger number of nuclear genes. Relationship among the remaining tribes differs in studies, but Chasmanthieae and Zeugiteae are being consistently revealed as sister lineages (Figure 3-1). Also often grouped together are Centotheceae, Cyperochloeae and Thysanolaeneae, although the exact relationship among these three tribes deserves further investigation. Positions for Gynerieae,

Tristachyideae and Steyermarkochloeae seems to be highly affected by sampling and molecular markers used and is basically unresolved. Notably, the monophyly of Steyermarkochloeae (*Arundoclaytonia* and *Steyermarkochloa*) is still questionable. *Arundoclaytonia* was originally included in this tribe based on morphology of spikelets (Davidse and Ellis, 1987), but molecular data showed it is closer to Chasmanthieae, while the placement of *Steyermarkochloa* is possibly sister to Tristachyideae but lacked support (Morrone et al., 2012). As a newly defined tribe, Jansenelleae is underrepresented in phylogenetic studies, but Bianconi et al. (2020) and Welker et al. (2020) both reported it as sister to (Andropogoneae, Arundinelleae) based on plastome and nuclear genes.

Relationship within Panicoid tribes has also been a focus of many studies. With over 1,200 species in ninety genera, **Andropogoneae** was further divided into eleven subtribes by Clayton and Renvoize (1986). Skendzic et al. (2007) sampled ten out of these eleven subtribes but found out most of them to be not monophyletic based on ITS and *trn*L-F sequences. They also failed to get consistent positions among different markers (ITS, *trn*L-F and combined) for *Arundinella*. Teerawatananon et al. (2011) performed combined analysis of (*trn*L–F + atpβ–*rbc*L + ITS) and confirmed the monophyly of subtribes Chionachninae, Coicinae, Dimeriinae, Germainiinae and Tripsacinae, but indicated that the subtribal classification of Clayton and Renvoize (1986) needs to be revised. Their results clearly supported *Arundinella* to be sister to the rest of Andropogoneae, while *Garnotia* (also in Arundinelleae according to Soreng et al. 2017) was nested among other Andropogoneae lineages. GPWG II (2012) reported a phylogeny based on three plastid genes where *Garnotia* and *Arundinella* forms a monophyletic tribe **Arundinelleae** sister to the rest of Andropogoneae. Similarly, Besnard et al. (2013) got an ITS-based result that supported this relationship. Arundinelleae was however treated as a subtribe (Arundinellinae) of Andropogoneae by Kellogg (2015), although the author acknowledged the sister relationship between these two lineages.

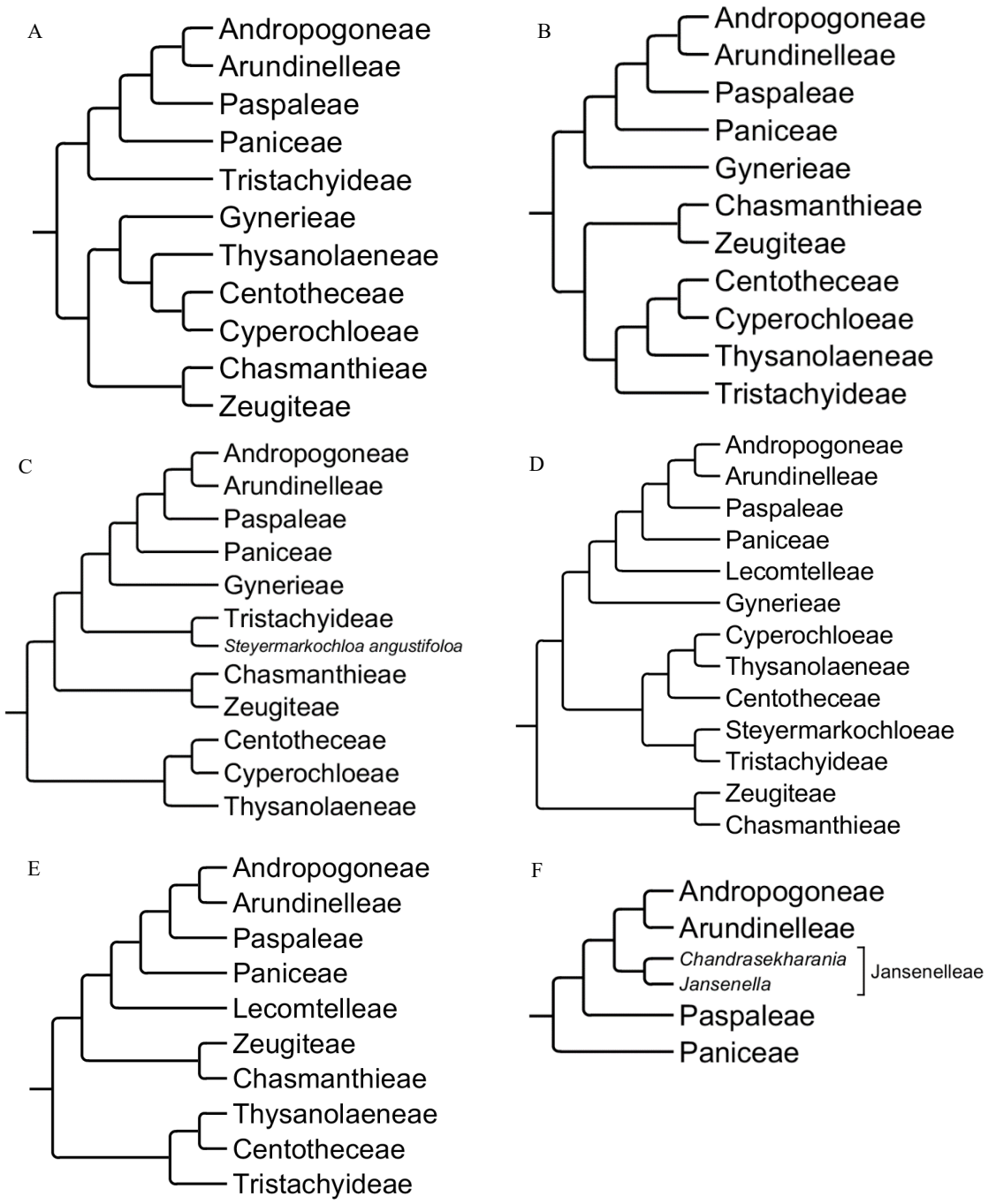


Figure 3-1: A comparison of Panicoideae phylogeny from previous studies. Name of tribes are shown. A: Sánchez-Ken et al. (2010), based on four genes, including one nuclear gene (GBSSI). B: GPWG II (2012), based on three plastid genes. C: Morrone et al. (2012), based on *ndhF*. D: Soreng et al. (2017), based on a plastid phylogeny as a backbone. E: Saarela et al. (2018), based on complete plastomes. F: Bianconi et al. (2020), based on plastome and nuclear genes.

Polyploidy is reported in several **Andropogoneae** species, including *Panicum*, *Saccharum*, and *Zea mays* (Hamoud et al., 1994; Wei et al., 2007; Zhang et al., 2019). To investigate the effect of polyploidy on diversification, Estep et al. (2014) reconstructed a phylogeny with one hundred Andropogoneae species using four low-copy nuclear loci and identified thirty-two polyploidy events based on gene tree topology. Although no evidence supported a positive correlation between polyploidy and diversification rate, their results indicated a reticulate evolutionary history of Andropogoneae. Nevertheless, the tribe as a whole is monophyletic in above-mentioned studies. Considering inconsistence among earlier studies, more-recent literatures tend to reduce the number of subtribes (seven in Kellogg 2015; nine in Soreng et al. 2017). However, as more sequencing data is available, Welker at al. (2020) proposed an updated Andropogoneae phylogeny with fourteen subtribes using complete plastome sequences, although leaving positions of several genera unresolved. Regardless of subtribal classifications, the foundation of phylogeny is species relationship, and the circumscription of Andropogoneae could be further improve as more comprehensive sampling is accomplished.

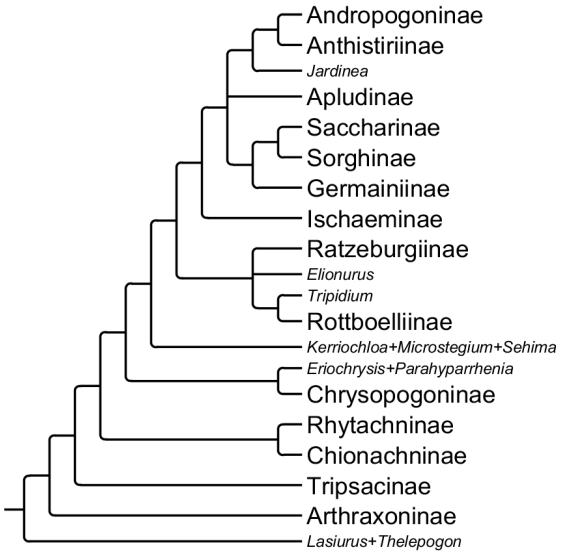


Figure 3-2: Relationship of subtribes in Andropogoneae, modified from Welker et al. (2020), a summary of their analyses based on complete plastome.

**Paspaleae** was referred to as the x=10 Paniceae (chromosome base number = 10) in earlier phylogenetic studies (Giussani et al., 2001; Aliscioni et al., 2003; Vicentini et al., 2008; Morrone et al., 2012), and this group is sister to Arundinelleae plus Andropogoneae. Three well supported subtribes, Arthropogoninae, Otachyriinae, and Paspalinae are recognized by Morrone et al. (2012), Kellogg (2015) and Soreng et al. (2017), and plastid gene-based studies supported Paspalinae to be sister to Arthropogoninae plus Otachyriinae (Aliscioni et al., 2012; Saarela et al., 2018; Bianconi et al., 2020). The monotypic genus *Reynaudia* was treated as uncertain by Morrone et al. (2012) due to lower support for its position and unknown chromosome number. Nevertheless, results from GPWG II (2012), Morrone et al. (2012), Welker et al. (2020) and Bianconi et al. (2020) consistently support *Reynaudia* to be basal to the rest of Paspaleae. To date, no nuclear gene-based phylogeny with a larger sampling size is available for this tribe, and the subtribal relationship deserves further investigation. Notably, all three subtribes contain both $C_3$ and $C_4$ species, while the basal lineage *Reynaudia* is $C_4$. Therefore, a reliable phylogeny is needed to study the evolution of photosynthetic pathway in this tribe.

**Paniceae** is comparable to Andropogoneae in term of species number (seventy-two genera, 1,254 species; Kellogg 2015) and is divided into seven subtribes by Soreng et al. (2017), although subtribe Dichantheliinae was not recognized and *Dichanthelium* was treated as *incertae sedis* by Kellogg (2015). In earlier literature (Giussani et al., 2001; Aliscioni et al., 2003; Vicentini et al., 2008), this tribe was named Paniceae x=9, to be distinguished from x=10 Paniceae which is now called Paspaleae. Morrone et al. (2012) proposed to split the obviously paraphyletic Paniceae s.l. (x=9 and x=10) into Paspaleae (x=10) and Paniceae s.s. (x=9), and further recognized six subtribes for Paniceae s.s.: Anthephorinae, Cenchrinae, Melinidinae, Boivinellinae, Neurachninae and Panicinae. The close relationship of Panicinae, Melinidinae and Cenchrinae were revealed by Morrone et al. (2012), GPWG II (2012) and Saarela et al. (2018), although the latter two studies supported a ((Melinidinae, Cenchrinae), Panicinae) topology and the former one supported

((Melinidinae, Panicinae), Cenchrinae). In all three studies, Neurachninae is sister to the above mentioned three subtribes, while Anthephorinae and Boivinellinae are at more basal positions. Notably, the position of *Dichanthelium* is inconsistent among these studies, but likely closer to the combined clade of Panicinae, Melinidinae and Cenchrinae. A couple of other *incertae sedis* genera, including *Homopholis*, *Sacciopelis* and *Trichanthecium*, all still pending for more support from molecular data to resolve their positions. For the large genus *Digitaria* with over 270 species, plastid phylogeny by GPWG II (2012) and Morrone et al. (2012) indicated it is likely not monophyletic, but more comprehensive sampling is needed to confirm this. Intriguingly, Paniceae is more complex in terms of photosynthetic pathway type, with all subtribes except Melinidinae composed of a mixture of $C_3$ and $C_4$ genera. Considering that the above-mentioned studies were largely based on plastid genes, a nuclear-gene phylogeny is needed to investigate the relationship among Paniceae subtribes.

In this project, I aimed to sample representative species for all available Panicoideae tribes and subtribes. For species that are easier to acquire, fresh material is preferred for RNA isolation followed by RNA-seq. Those with only herbarium samples will be subject to DNA isolation and genome sequencing. The goal is to get a well-resolved Panicoideae phylogeny using low-copy nuclear genes. With this phylogeny, the evolution of photosynthetic pathway is discussed, and an expanded *ppc* gene family analyses that include more Panicoideae species is performed.

### 3.1.2 GC content of genomes and genes

GC-content is the percentage of nitrogenous bases that are either guanine (G) or cytosine (C) for a certain fragment of DNA/RNA or for an entire genome, calculated as: (G+C) / (G+C+A+T) × 100%. GC content varies among genomes (among different species), within genome (among

different chromosomes /regions on chromosomes), among genes (or paralogs/gene copies), and within regions of a single coding sequence.

GC content of genome varies in angiosperms (see Table 3-1 for examples), with a rough range of 30 to 50%. Serres-Giardi et al. (2012) examined the GC content at third codon positions of coding sequences (GC3) across a wide range of plant, but no obvious pattern is observed in angiosperms. Nevertheless, Poaceae species generally have a higher genome GC content (usually > 40%), and a dichotomy of GC3 is observed in *Brachypodium distachyon*, *Sorghum bicolor*, *Oryza sativa* and *Zea mays* (Serres-Giardi et al., 2012, supplemetary data). To be more specific, in these grass species, the distribution of GC3 for genes fits better to a bimodal shape rather than unimodal, as in most other angiosperm species.

**Table 3-1: GC content of plant genomes**

| family | species | genome GC content |
|---|---|---|
| Asteraceae | *Helianthus annuus* (sunflower, GCA_002127325.2) | 38.8% |
| Solanaceae | *Solanum lycopersicum* (tomato, GCA_000188115.4) | 35.7% |
| | *Solanum tuberosum* (potato, GCA_000226075.1) | 35.6% |
| Brassicaceae | *Arabidopsis thaliana* (GCA_000001735.2) | 36.1% |
| Orchidaceae | *Phalaenopsis aphrodite* (moth orchid, GCA_003013225.1) | 30.4% |
| Rosaceae | *Malus domestica* (apple, GCA_002114115.1) | 39.3% |
| Musaceae | *Prunus persica* (peach, GCA_000346465.2) | 37.7% |
| Fabaceae | *Glycine max* (soybean, GCA_000004515.5) | 35.1% |
| Cyperaceae | *Carex parvula* (GCA_025461045.1) | 34.1% |
| | *Musa acuminata* (GCA_000313855.2) | 40.7% |
| Poaceae | *Oryza sativa* (GCA_001433935.1) | 43.6% |
| | *Zea mays* (GCA_902167145.1) | 46.8% |
| | *Triticum aestivum* (GCA_018294505.1) | 46.2% |
| | *Eleusine coracana* (GCA_021604985.1) | 39.6% |
| | *Phyllostachys edulis* (GCA_011038535.1) | 41.6% |

**Table 3-1: continued**

| | |
|---|---|
| *Pharus latifolius* (GCA_019359835.1) | 44.3% |
| *Streptochaeta angustifolia* (GCA_020804685.1) | 43.5% |

A well-known hypothesis about GC content evolution is that GC-rich DNA/RNA molecules is an adaptation to higher temperature, especially for prokaryotes, because GC pair with three hydrogen bonds is thermally more stable than AT/AU pair with two hydrogen bonds. However, both base paring between complementary DNA strands and stacking between adjacent bases contribute to the stability of DNA double helix, and the later one is found to be the main stabilizing factor (Yakovchuk et al., 2006). By comparative analyses among prokaryote genomes, Hurst and Merchant (2001) found no correlation between genomic GC content/GC content at 3$^{rd}$ codon positions and optimal growth temperature; on the other hand, their results showed that GC content of structural RNA is correlated with temperature.

GC content variation among paralogous genes could reflect codon usage bias, in which certain synonymous codons are preferred over others. Codon bias has been implicated as one of the major factors contributing to transcription rate, translation efficiency and mRNA stability. Hia et al. (2019) showed that in humans RNA binding proteins regulate mRNA half-life, depending on GC content and codon usage. RNAs with shorter half-lives were associated with AT3 codons, while those with longer half-lives were associated with GC3 codons. Newman et al. (2016) reported that codon bias and GC content contributes to the different expression levels of TLR7 and TLR9 (genes related to immunity in human), and that the major factor causing the difference is transcription rate. They proposed that suboptimal codon bias, which correlates with lower guanine-cytosine (GC) content, limits transcription of certain genes. Using three variants of *URA3* gene with different GC content, Kiktev et al. (2018) demonstrated that GC content has a positive correlation with mutation and recombination rates in yeast.

The above-mentioned studies focused on effect of GC content variation on the biochemical properties of mRNA for specific genes, but few is known about how GC content is different among paralogs of the same gene family. One relevant study was by Bowers et al. (2022), they examined the genomes of five dicots and four grasses and reported that syntenic genes have significantly higher 3$^{rd}$ codon GC content compared to non-syntenic genes, especially at the 5' end. Since synteny is often used as a proxy for gene duplication, the authors proposed that the syntenic genes in grasses have undergone fewer duplications or the duplicates were purged by selection. Whether this statistically significant conclusion applies to gene families that are involved in photosynthesis deserves further investigation. The *ppc* gene family (introduced in 2.1.4) includes five to six paralogs of highly similar sequences, but most functionally $C_4$ *ppc* genes are from the *ppc-B* clade. The $C_4$ version of enzyme PEPC participates in the initial assimilation of $CO_2$ in mesophyll cells, whereas non-$C_4$ PEPC is involved in other pathways and not preferentially expressed in leaves, so whether GC content is related to the expression pattern of *ppc* genes needs to be investigated.

## 3.2 Methods

*Some of the methods used for this chapter are the same as described in chapter 2 and are not shown here to reduce redundancy. These includes but are not limited to: grass sample collection and sequencing (2.2.1), trimming of raw sequencing data (2.2.2), sequence alignment and reconstruction of single-gene ML trees (2.2.4), and coalescent analyses by Astral (2.2.6).

### 3.2.1 Assembly of genome skimming and transcriptomic data

For shotgun genome sequencing data (at relatively shallow depth, so-called genome skimming), Trimmomatic 0.36 (Bolger et al., 2014) was also implemented to remove sequencing adaptors and low-quality regions. Trimmed genomic data sets were assembled to contig level by SOAPdenovo2 (2.04-r240) (Luo et al., 2012). For transcriptomic data, pair-end sequencing data sets were first trimmed by Trimmomatic (Trinity 2.2.0 plug-in) using default settings. Transcriptome assembly was done by Trinity (V 2.2.0) (Grabherr et al., 2011) with default parameters on Penn State Roar server. Deduplication of assembly contigs were done by CD-HIT-EST (V 4.6.8) (Fu et al., 2012). Coding sequences were extracted from deduplicated contigs by TransDecoder (V 5.3.0). Statistics on genomic contigs (for genome skimming data) and none-redundant coding sequences (for transcriptomic data) were calculated by statswrapper.sh (a bbmap tool, V 38.33) to check assembly quality.

### 3.2.2 Obtaining target genes from genome skimming and transcriptome data

The genome/transcriptome sequences of ten Poaceae species (*Brachypodium distachyon*, *Eleusine coracana*, *Hordeum vulgare*, *Lygeum spartum*, *Oryza sativa*, *Phaenosperma globosum*, *Phyllostachys heterocycle*, *Setaria italica*, *Sorghum bicolor* and *Stipa aliena*) were used to compare

and identify putative low-copy (one or two copies per species) nuclear genes across Poaceae family by OrthoMCL v1.4 (Li et al., 2003). The HMM files of 1,234 OGs (orthologous groups) identified were used as the seeds for HaMStR (13.2.6) (Ebersberger et al., 2009) to search and obtain the corresponding putative orthologous sequences from the genomic contigs and none-redundant coding sequences datasets. Cutoff e-values for blast and hmm search were both set to 1e-20. There is only one sequence retained per dataset for each seed gene (each orthologous, or OG), and in cases there are fragments they are combined to represent the whole sequence.

### 3.2.3 Purging potential paralogs from single-gene trees

For manual inspection of single-gene trees, all the tree terminals are labelled with a string showing its taxonomic information from tribe name all the way to species name (e.g., Andropogoneae-Saccharinae-*Sorghum bicolor*), and single-gene trees are viewed in FigTree v1.4.4 one by one, manually checked for questionable branches. Although subjective decision cannot be fully excluded, a criterion was set that any tree with more than ten sequences grouped together and formed a long branch/in an obviously questionable position is marked as a “bad tree” containing potential paralogs. Following the manual inspection, 250 single-gene trees were marked as “bad trees” and excluded from downstream coalescent analyses. The remaining 984 single-gene tree were subjected to a second round of inspection by TreeShrink1.3.9 (Mai and Mirarab, 2018) to further remove long branches that remained in the single-gene trees (these long branches did not reach the ten-sequence threshold). Single-gene trees and FASTA alignment files of the 984 genes were used as input for TreeShrink, and “all-genes” option was used, with “-b 3 -k 50”. The outgroup species for Panicoideae were protected from deleting. The sequences corresponding to long branches were deleted from output FASTA alignment files, and the FASTA files were each re-aligned to improve quality.

### 3.2.4 Inspecting coalescent analysis behavior by mock data

To inspect the behavior of coalescent analysis (as implemented in Astral) in regard to the placement of rogue taxa, single-gene trees were made up by manually specifying the relationship among twelve species of fixed relationships plus a species "X" that was *incertae sedis* (or a so-called rogue taxon). In other words, the mock data starts from gene trees directly, instead of actual sequences of genes, because I focused on the step from gene trees to coalescent trees and wanted to exclude variables induced from other steps. These gene trees represent all possible placements of species X and are named accordingly (see Figure 3-13 for examples). Coalescent trees were reconstructed by Astral 5.7.8 from seven sets of gene trees with the following compositions (H*-C* represents copies of the same gene tree, just as H*-copy*. For example, HABCDE-C1 and HABCDE-C2 are the same as HABCDE):

Set 1:

H0, HA, HA1, HA1A2, HA2, HA3, HA3A4A5, HA4, HA4A5, HA5, HAB, HABC, HABCD, HABCDE, HABCDE-C1, HABCDE-C2, HABCDE-C3, HB, HB1, HB2, HB2B3, HB3, HC1, HD, HD1, HD2, HE1.

Set 2:

H0, HA, HA1, HA1A2, HA2, HA3, HA3A4A5, HA4, HA4A5, HA5, HAB, HABC, HABCD, HABCDE, HB, HB1, HB2, HB2B3, HB3, HC1, HD, HD1, HD2, HE1.

Set 3:

H0, HA, HA1, HA1A2, HA1A2-C1, HA2, HA2-C1, HA3, HA3A4A5, HA4, HA4A5, HA5, HAB, HABC, HABCD, HABCDE, HB, HB1, HB2, HB2B3, HB2B3-C1, HB3, HC1, HD, HD1, HD2, HE1.

Set 4:

H0, HA-C1, HA, HA-C1, HA1, HA1A2-C1, HA1A2, HA2-C1, HA2, HA3-C1, HA3, HA3A4A5-C1, HA3A4A5, HA4-C1, HA4, HA4A5-C1, HA4A5, HA5-C1, HA5, HAB, HABC, HABCD, HABCDE, HB-C1, HB, HB1-C1, HB1, HB2-C1, HB2, HB2B3-C1, HB2B3, HB3-C1, HB3, HC1, HD, HD1, HD2, HE1.

Set 5:

H0, HA-C1, HA, HA1-C1, HA1, HA1A2-C1, HA1A2, HA2-C1, HA2, HA3-C1, HA3, HA3A4A5-C1, HA3A4A5, HA4-C1, HA4, HA4A5-C1, HA4A5, HA5-C1, HA5, HAB, HABC, HABCD, HABCDE, HB-C1, HB-C2, HB-C3, HB, HB1-C1, HB1-C2, HB1-C3, HB1, HB2-C1, HB2-C2, HB2-C3, HB2, HB2B3-C1, HB2B3-C2, HB2B3-C3, HB2B3, HB3-C1, HB3-C2, HB3-C3, HB3, HC1, HD, HD1, HD2, HE1.

Set 6:

H0, HA-C1, HA, HA1-C1, HA1, HA1A2-C1, HA1A2, HA2-C1, HA2, HA3-C1, HA3, HA3A4A5-C1, HA3A4A5, HA4-C1, HA4, HA4A5-C1, HA4A5, HA5-C1, HA5, HAB, HABC, HABCD, HABCDE, HB-C1, HB-C2, HB-C3, HB-C4, HB, HB1-C, HB1-C2, HB1-C3, HB1-C4, HB1, HB2-C1, HB2-C2, HB2-C3, HB2-C4, HB2, HB2B3-C1, HB2B3-C2, HB2B3-C3, HB2B3-C4, HB2B3, HB3-C1, HB3-C2, HB3-C3, HB3-C4, HB3, HC1, HD, HD1, HD2, HE1.

Set 7:

H0, HA, HA1, HA1A2, HA2, HA3, HA3A4A5, HA4, HA4A5, HA5, HAB, HABC - Copy, HABC, HABCD-C1, HABCD, HABCDE-C1, HABCDE, HB, HB1, HB2, HB2B3, HB3, HC1-C1, HC1, HD-C1, HD, HD1-C1, HD1, HD2-C1, HD2, HE-C1, HE1.

### 3.2.5 *ppc* gene family analysis

To obtain putative *ppc* transcripts from transcriptome datasets, amino acid sequences of PEPC from *Zea mays* (Panicoideae), *Streptochaeta spicata* (Anomochlooideae) and *Cyrtococcum patens* (Panicoideae) were used as queries to perform tblastn against the coding sequences from Poaceae and outgroup species. All Poaceae subfamilies were included, except Puelioideae for which only genome skimming data is available and no *ppc* sequences of good quality could be obtained. Representative species were selected from each subfamily based on assembly quality and $C_3$/$C_4$ status; for Panicoideae, all species with transcriptome data were included. After blastn search, the corresponding nucleotide coding sequences were obtained, and sequences shorter than 600bp (200 aa) were excluded. Additional *ppc* coding sequences from genomes were downloaded from Phytozome 13 according to annotation of genes followed by manual inspection. Putative ppc coding sequences were translated into amino acid sequences, aligned by MAFFT, and corresponding nucleotide alignment was generated. Preliminary analyses were done to identify bacterial-type PEPC (BTPC) by sequence features and branch length on gene trees.

## 3.3 Results

### 3.3.1 Panicoideae phylogeny based on low-copy nuclear genes

After preliminary analyses, a final set of 285 samples were included for Panicoideae phylogeny, including 136 genomic datasets (13 from public genomes and 123 de novo assembled) and 149 transcriptomic datasets. To reduce the effect of low-quality sequences, starting from the 1234-gene set, 250 single-gene trees were purged due to higher percentage of potential paralogs (method is described in 3.2.3). The remaining 984 gene set were further filtered based on species coverage, resulting in three subsets of 809 (70%), 601(80%) and 439(85%) genes, respectively. Coalescent analyses were performed on each set of the gene trees, and comparison is made and summarized in figures 3-3. Classification of Panicoideae used in this chapter follows Soreng et al. (2017), with updates from Welker et al. (2020).

Our coalescent analyses based on four sets of nuclear genes consistently support the same relationship among eleven Panicoideae tribes sampled in this project (Figure 3-3). *Dichaetaria wightii*, treated as *incertae sedis* of Panicoideae (Soreng et al., 2017), is clearly the first diverging lineage of the whole subfamily. This species was put in Arundinoideae by Kellogg (2015), but our phylogeny with representatives from all PACMAD subfamilies confirmed its proximity to Panicoideae. Chasmanthieae and Zeugiteae forms the next diverging clade, followed by Thysanolaeneae plus Centotheceae. *Arundoclaytonia*, once and still treated as a Steyermarkochloeae genus by Soreng et al. (2017), is revealed to be sister to Zeugiteae, a position similar to some previous studies (Sa'nchez-Ken and Clark 2007; Morrone et al. 2012). Notably, Kellogg (2015) combined Zeugiteae and Chasmanthieae into Chasmanthieae and included *Arundoclaytonia*, a classification more reasonable based on my results. The next diverged lineage is the monotypic tribe Gynerieae with *Gynerium sagittatum*. These basal tribes contain only $C_3$ taxa,

while the remaining tribes (Tristachyideae, Paniceae, Paspaleae, Arundinelleae and Andropogoneae) are a mixture of $C_3$ and $C_4$ grasses.

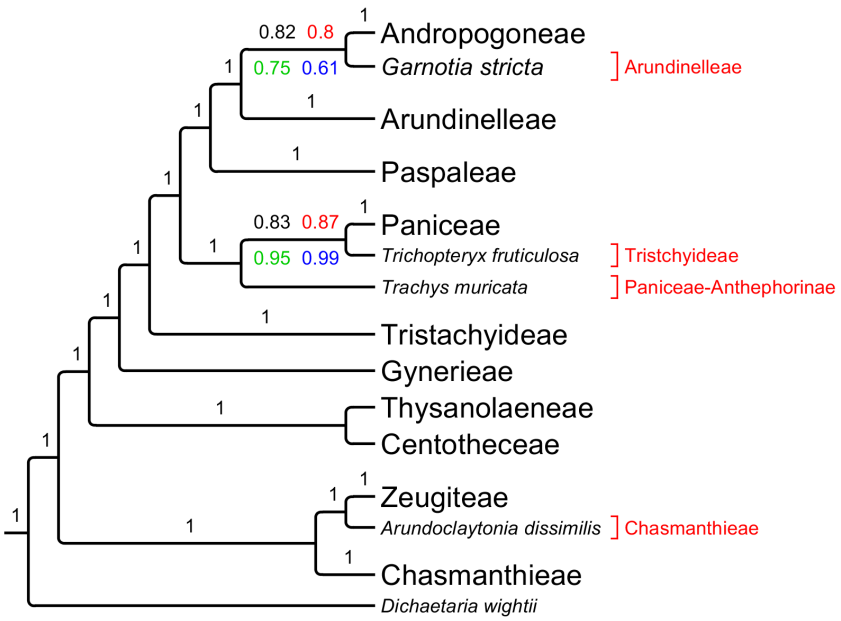


Figure 3-3: A summary of Panicoideae phylogeny based on four coalescent analyses. Tribe names are in normal font, species that make the corresponding tribes not monophyletic are italicized and labelled to the right. Posterior probability support values are labelled above branches. For branches received 1.0 from all four coalescent analyses, only a single "1" is labelled. For the two branches with different support values, numbers are in different colors: black, 984-gene; red, 809-gene; green, 601-gene; blue, 439-gene.

**Tribe Andropogoneae** as a whole is consistently revealed as monophyletic in all coalescent analyses (Figure 3-4), although some of the subtribes are split into multiple lineages. Results clearly support the monotypic subtribe Arthraxoninae (*Arthraxon*) as the first diverging lineage in Andropogoneae, followed by a clade containing two sub-clades, one with Tripsacinae (*Zea* and *Tripsacum*) and the other with six genera put in subtribe Ratzeburgiinae by Welker et al. (2020): *Hemarthria*, *Heteropholis*, *Hackelochloa*, *Mnesithea*, *Eremochloa*, and *Thaumastochloa*, plus *Phacelurus* which was placed *incertae sedis* by their analyses. The above-mentioned seven genera were all classified as in subtribe Rottboelliinae by Soreng et al. (2017), although the authors indicated that this subtribe is probably paraphyletic. Welker et al. (2020) revised Rottboelliinae to include only *Chasmopodium*, *Coix* and *Rottboellia*, and moved most of other genera previously in

this subtribe to Ratzeburgiinae. Our results support the monophyly of Ratzeburgiinae, suggesting *Phacelurus* should be included, but differences emerged for the relationship among Ratzeburgiinae,

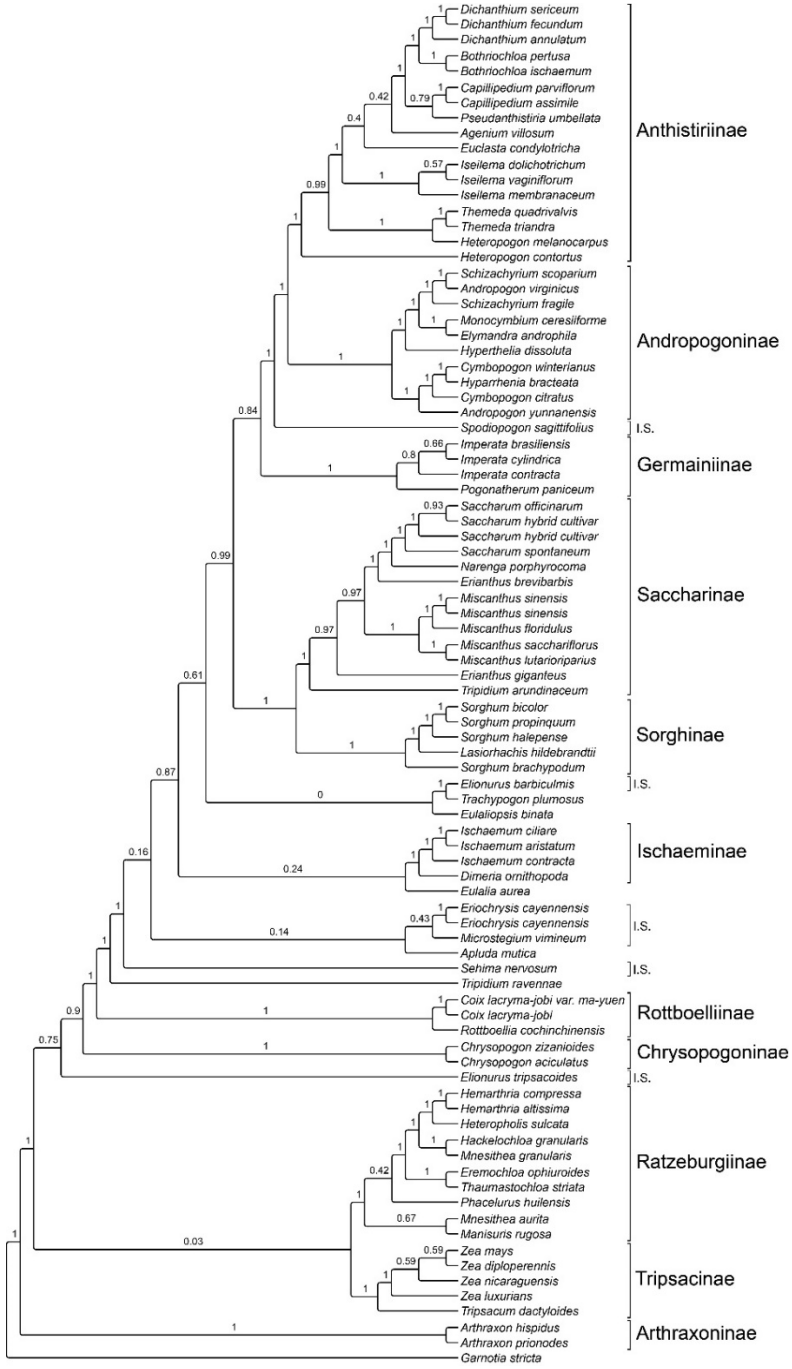


Figure 3-4: Andropogoneae phylogeny from 984-gene coalescent analysis. Names of subtribes are labelled to the right; posterior probability support values are shown above the branches. *Garnotia stricta* (Arundinelleae) is shown as an outgroup.

Tripsacinae and Rottboelliinae. In all coalescent trees, Tripsacinae is sister to Ratzeburgiinae, with weak but increasing support values as reduce the number of genes is reduced for coalescent analyses (PP=0.03, 0.37, 0.69 and 0.97 for four gene sets, respectively), and Rottboelliinae is inside the clade containing the rest of Andropogoneae. The position of *Elionurus* is unclear from the major coalescent trees, but a supplementary analysis with reduced taxa (not shown) suggested it might be basal to other Ratzeburgiinae sampled in our project. Genus *Chrysopogon* forms a separate lineage, followed by Rottboelliinae (*Coix* and *Rottboellia*). *Agenium* was put in Andropogoninae by Kellogg (2015), but in Saccharinae by Soreng et al. (2017). My results show it is sister to a clade containing *Dichanthium*, *Bothriochloa*, *Capillipedium* and *Pseudanthistiria*, named subtribe Anthistiriinae by Welker et al. (2020). For the position of *Euclasta*, my results don't agree with either classification of Soreng et al. (2017) or phylogeny by Welker et al. (2020). The reason is unknown but probably due to differences in sampling. *Heteropogon*, *Iseilema* and *Themeda* were also included in Anthistiriinae by Welker et al. (2020), and in my results they are closely grouped together with the clade mentioned just above, supporting the monophyly of Anthistiriinae, although *Heteropogon* may be paraphyletic. *Cymbopogon*, although placed in Anthistiriinae by Welker et al. (2020), is maximumly supported to be grouped together with *Schizachyrium*, *Andropogon*, *Monocymbium*, *Elymandra*, *Hyperthelia* and *Hyparrhenia*, all of which are Andropogoninae. Among the genera just mentioned, *Schizachyrium*, *Andropogon* and *Cymbopogon* are not monophyletic in my nuclear phylogeny. *Spodiopogon*, which was treated as *incertae sedis* by Welker et al. (2020), is clearly (PP=1.0 in all coalescent trees) sister to the combined clade of Anthistiriinae and Andropogoninae. Our sampling covered *Imperata* and *Pogonatherum* for Germainiinae, and these two genera forms a monophyletic clade sister to Anthistiriinae plus Andropogoninae. *Sorghum* plus *Lasiorhachis*, both in Sorghinae, are sister to Saccharinae (*Saccharum, Narenga=Miscanthus, Erianthus=Saccharum, Miscanthus*) plus *Tripidium arundinaceum* (*incertae sedis* by Welker et al., 2020). The other *Tripidium* species included, *Tripidium ravennae*, is however at a more basal

position, making this genus not monophyletic, although *Tripidium* is reported to be monophyletic by Evans et al. (2019) and Welker et al. (2020).

Our sampling covers *Arundinella* and *Garnotia* for **tribe Arundinelleae**. Different from previous studies (Teerawatananon et al., 2011; GPWG II, 2012; Besnard et al., 2013), our results revealed a new relationship where *Arundinella* is sister to *Garnotia* plus the entire Andropogoneae, although a slightly decreasing trend is observed for support values of the node leading to (*Garnotia*, Andropogoneae), as the number of gene trees is reduced for coalescent analyses (Figure 3-3). However, quartet value (q1) is relatively low (0.37 for 984-gene coalescent tree) for this node, suggesting substantial conflicts among single-gene trees. Therefore, although *Arundinella* is clearly monophyletic, the exact position of *Garnotia* awaits further verification.

**In tribe Paspaleae,** *Reynaudia filiformis* is revealed to be the first diverging lineage (Figure 3-5), consistent with phylogenies from GPWG II (2012), Welker et al. (2020) and Bianconi et al. (2020). *Anthaenantia*, previously put in Otachyriinae by Kellogg (2015) but in Paspalinae by Soreng et al. (2017), is clearly sister to the other Otachyriinae genera (*Hymenachne*, *Steinchisma* and *Otachyrium*) sampled in this project, which form a monophyletic clade in all coalescent trees. For Arthropogoninae, phylogeny based on three plastid genes by GPWG II (2012) supported its monophyly, while Morrone et al. (2012) received weak support for it with only *ndhF* sequences. In my results, Arthropogoninae is paraphyletic in all coalescent analyses. *Arthropogon* and *Homolepis* together are sister to Otachyriinae, while *Oncorachis* and the remaining genera form two lineages that are successively sister to subtribe Paspalinae. On the other hand, Paspalinae itself and the sampled genera it contains are all monophyletic. One unexpected result is that *Chaetium festucoides*, which is the type species of its genus and was classified into Paniceae–Melinidinae (according to Soreng et al. 2017), is revealed to be nested in Paspalinae, sister to genus *Axonopus*.

*Chaetium festucoides* was never included in a molecular phylogeny before, so this relationship deserves further investigation.

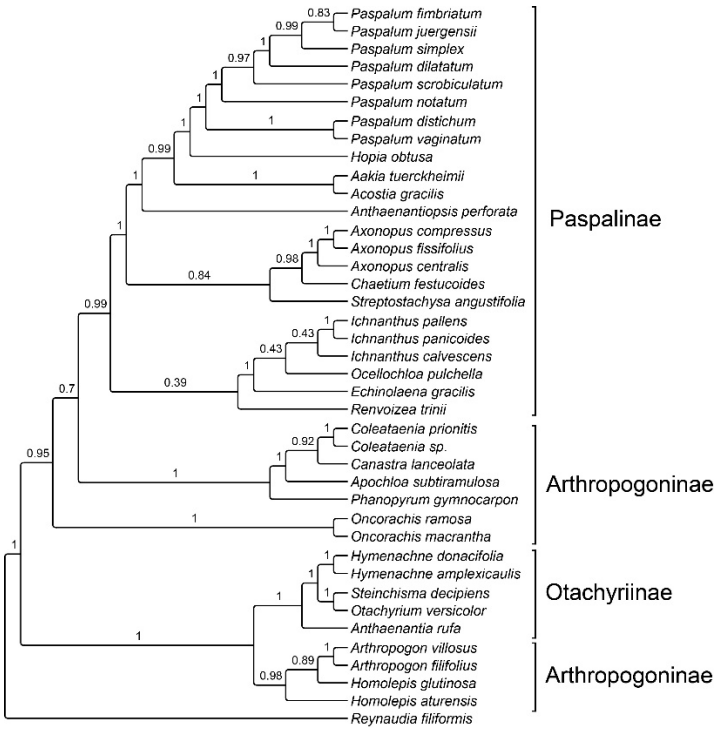


Figure 3-5: Paspaleae phylogeny from 984-gene coalescent analysis. Names of subtribes are labelled to the right; posterior probability support values are shown above the branches.

For **Paniceae**, coalescent analyses based on four sets of genes consistently support its monophyly (Figure 3-6), but with *Trichopteryx fruticulosa* embedded as the second diverging branch followed by *Trachys muricata* (position for this species is further investigated in 3.3.2). *Trichopteryx* was put in tribe Tristachyideae by Kellogg (2015) and Soreng et al. (2017), both classifications were based on previous literature and this genus in fact lacked support from molecular data. The other species in this genus, *Trichopteryx elegantula*, is clearly a member of Tristachyideae based on my results, so the monophyly of *Trichopteryx* is questionable. The position of *Alloteropsis* and *Dichanthelium* differs among the four coalescent trees, with two alternative topologies. Nevertheless, these two genera together with Boivinellinae compose a clade that is

sister to the rest of Paniceae. The relationship among Cenchrinae, Melinidinae, *Panicum*+*Louisiella*, and Anthephorinae is pretty clear and consistent among results. Notably, *Stereochlaena cameronii*, formerly placed in Cenchrinae based on morphological data by Morrone et al. (2012), is shown to be in Anthephorinae with high support values. *Homopholis* is clearly monophyletic in my results, and sister to a clade of Neurachninae, Anthephorinae, Paniceae, Melinidinae and Cenchrinae combined.

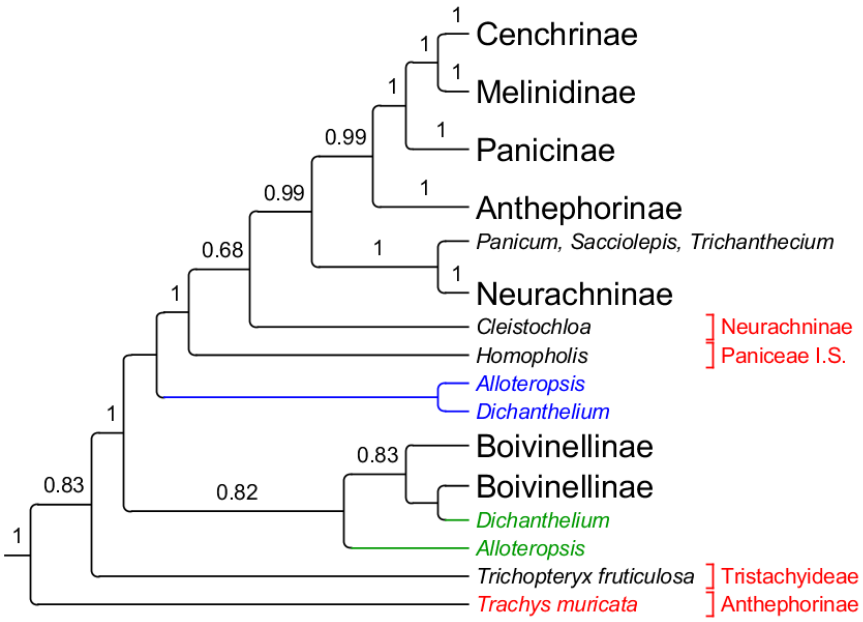


Figure 3-6: Relationship of Paniceae subtribes based on 984-gene coalescent tree using the final sample set. For *Alloteropsis* and *Dichanthelium* (branches in green), the alternative position (branches in blue) is also shown.

### 3.3.2 Improvement of Panicoideae phylogeny by purging long branches and using reduced datasets

While the Panicoideae phylogeny is overall well-resolved by low-copy nuclear genes, preliminary coalescent analyses revealed unexpected positions for several species, including *Polytoca digitata*, which was put in Chionachne by previous studies (Soreng et al., 2017; Welker

et al., 2020). In the initial coalescent tree based on 1234 nuclear genes, this species is however sister to Paspaleae, Arundinelleae and Andropogoneae combined, with high support values on the associated nodes (Figure 3-7, left). Considering the gene sequences for this species is from a genome-skimming dataset with relatively low quality (248/1234 genes were found), I checked all the single-gene trees for the position of this specific species and found no dominant topology (40 gene trees placed it in tribe Paniceae, 36 placed it outside Panicoideae, 61 placed it in Andropogoneae, but exact positions varied; other data not shown). However, during manual inspection, long branches associated with this species is also frequently observed (see examples in Figure 3-8), questioning the validity of its position in single-gene trees, because long branches is often an indication of potential paralogs or low-quality sequences.

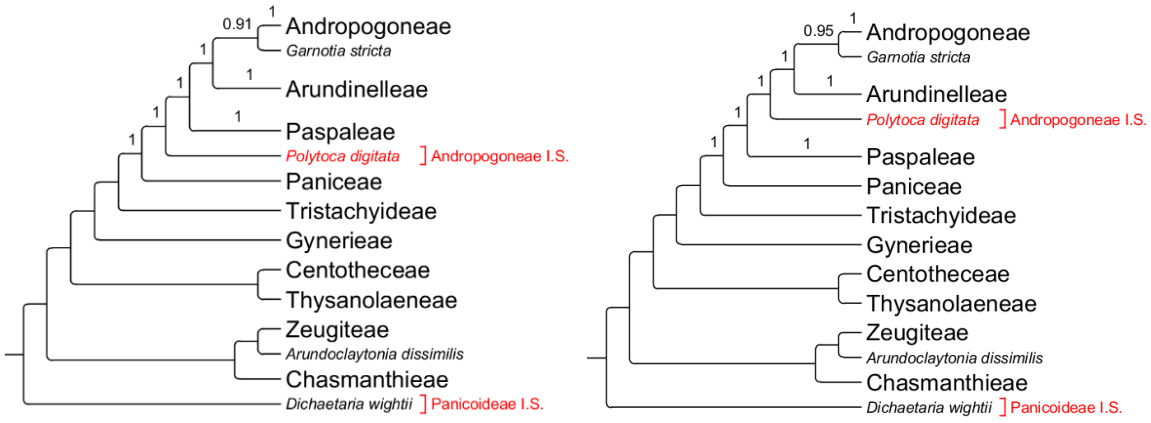


Figure 3-7: Removal of long branches in single-gene trees improved Panicoideae phylogeny. Left: Panicoideae phylogeny from 1234 original single-gene trees. Right: 1234-gene coalescent tree after removing long branches from single-gene trees. Values above branches are posterior probability values calculated by Astral. Branches without marks received the highest support (PP=1.0).

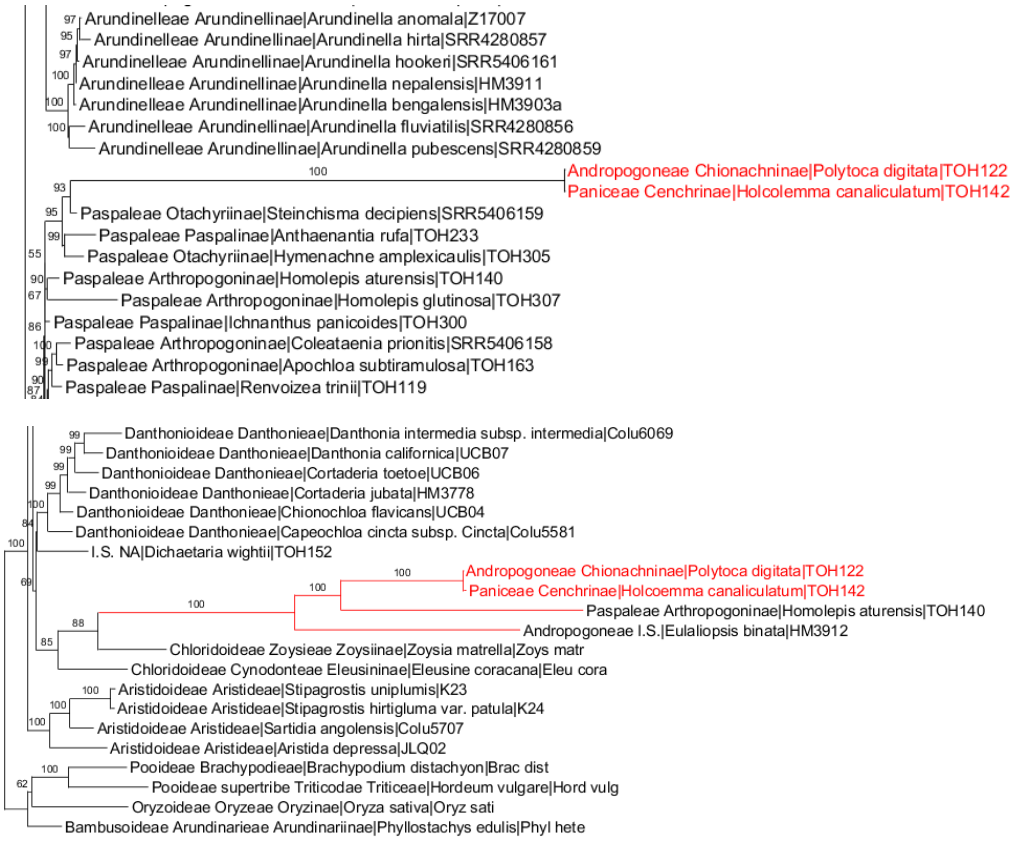


Figure 3-8: Examples of long branches in single-gene trees. Part of gene tree OGs_8128 (top) and OGs_10500 (bottom) are shown, and long branches are marked by red. Horizontal distance between two tips (species) represent the evolutionary changes observed. Vertical distance does not have significance for interpretating evolution.

TreeShrink 1.3.9 (Mai and Mirarab, 2018) was used to purge long branches from the all the 1234 single-gene trees out of which 248 contain *Polytoca digitata*. Seventy-three single-gene trees were identified as contain a long branch for *Polytoca digitata*, and this species was deleted from the corresponding single-gene trees as well as the alignment of corresponding genes. As a comparison, for *Dichaetaria wightii* which is found in 970 single-gene trees, only nine were detected as contain a long branch for this species. A coalescent tree using the same 1234 single-gene trees but with long branches deleted shows improvement for the position of *Polytoca digitata*, moving it to a position closer to Andropogoneae (Figure 3-7, right).

A more significant improvement is observed for the position of *Holcolemma canaliculatum.* This species was supposed to be in subtribe Cenchrinae (Soreng et al., 2017), but in the initial 1234-gene coalescent tree it was at a very basal position of the whole Paniceae tribe (Figure 3-9, left), and support values along the backbone nodes are high. After TreeShrink (sequences of this species was removed for 274 out of 640 single-gene trees), it is in a position much closer to Cenchrinae, sister to Melinidinae (Figure 3-9, right). Indeed, during manual inspection I noticed this species is even more frequently associated with long branches compared to *Polytoca digitata*. Also, more genes remained (366) for this species after purging of long branches, resulting in a better resolved position.

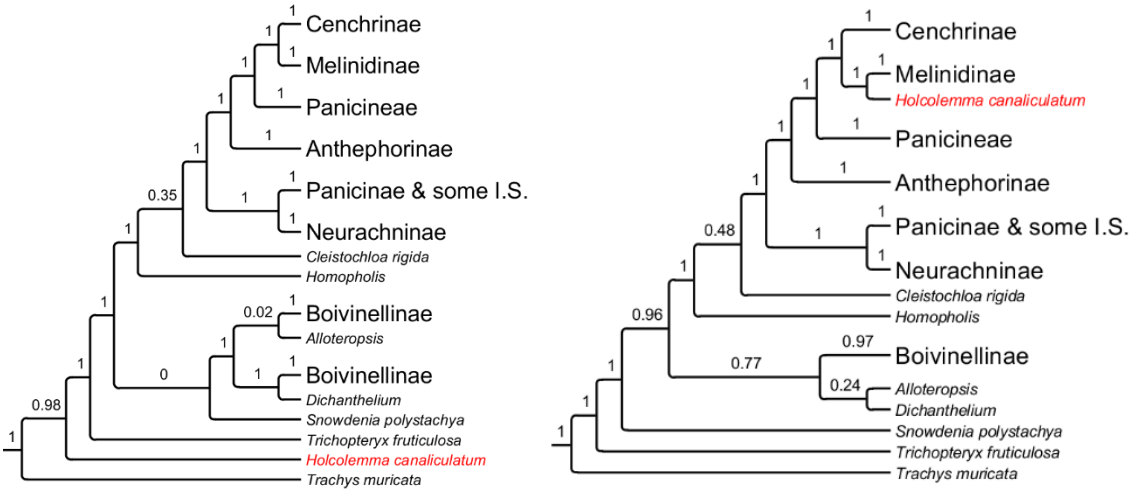


Figure 3-9: Summary of relationships among subtribes and some genera in tribe Paniceae, from original 1234-gene coalescent tree (left) and from 1234-gene coalescent tree post-TreeShrink (right). *Holcolemma canaliculatum* is marked in red to show the drastic change of its position.

The examples from *Polytoca digitata* and *Holcolemma canaliculatum* manifested that removing long branches from single-gene trees could improve the position of certain species in Panicoideae phylogeny. Based on the information I gathered from manually inspection of single-gene trees, I hypothesized that coalescent analyses tend to place a species at a basal position of a clade when there is considerable amount of conflict among single-gene trees for it, but the conflicts are confined within this clade (see more analyses in 3.3.3).

However, a couple of species remained in questionable positions in the post-TreeShrink coalescent results. For example, *Trachys muricata* is supposed to be in subtribe Anthephorinae, but in the 234-gene coalescent tree post-TreeShrink it is still at the first diverging branch in Paniceae. After a careful inspection of some post-TreeShrink single-gene trees I noticed that when a clade of over ten sequences as a whole form a long branch it is usually not deleted, probably due to the fact that the portion is above the threshold of TreeShrink to be considered for deletion. Nevertheless, I noticed that in some single-gene trees there are long-branched clades composed of more than ten, sometimes even more than twenty sequences, at questionable positions. Therefore, all the original 1,234 single-gene trees were manually inspected and those with more than ten sequences clustered as long branch(es) in odd positions were marked as bad trees and excluded from the 1,234 set. A new coalescent tree was generated on the remaining 984 single-gene trees which also went through TreeShrink to remove long branches (results already showed in 3.3.1).

As expected, the new 984-gene coalescent tree showed further improvement for Panicoideae phylogeny. Specifically, *Trachys muricata* was now clustered with other Anthephorinae species (Figure 3-10). The support value (PP) for this subtribe is, however, a zero. Considering that in the 1234-gene coalescent tree prior to TreeShrink the value for the branch of Anthephorinae is 1.0, *Trachys muricata* is suspected to cause the problem, because it is the only one whose position was changed among members of this subtribe. Therefore, I decided to manually check all these 193 single-gene trees (out of 984) that contain this species and put them into categories. *Trachys muricata* was placed among Andropogoneae, Arundinelleae or Paspaleae in fifty-eight single-gene trees. 126 single-gene trees placed it in Paniceae, among which sixty-four placed it in Anthephorinae, but the exact position varied. Nine single-gene trees placed it in other Panicoideae tribes or outside the subfamily. Therefore, as the largest category, around one third (64/193) of the single-gene trees placed it in the correct subtribe, but this percentage is probably not high enough for coalescent method to report a decent support value for its position. It is

reasonable to place *Trachys muricata* in Anthephorinae, but its precise position needs to be further investigated.

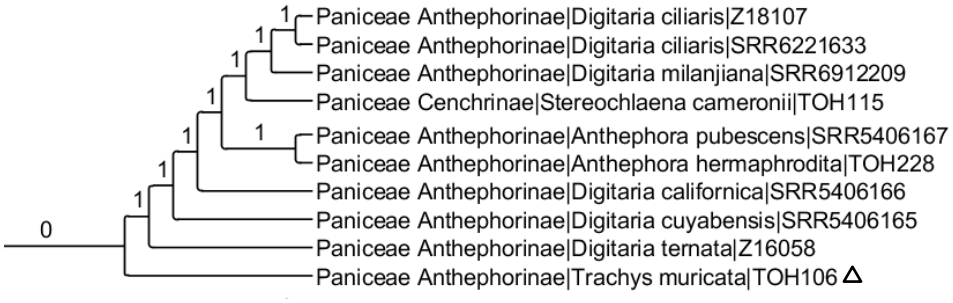


Figure 3-10: Part of a 984-gene coalescent tree, showing subtribe Anthephorinae. This was an earlier tree before taxa were finalized. Note the zero value at the branch leading to Anthephorinae. *Trachys muricata* is marked by a triangle.

Single gene trees that placed this species in other tribes might be a result of low sequence quality/shorter sequence length, considering their smaller portion. A coalescent tree using only the 64 gene trees that put *Trachys muricata* in Anthephorinae was reconstructed to get a resolution for the position of *Trachys muricata* inside this subtribe. In the 64-gene coalescent tree (Figure 3-11), the PP value for the stem node of Anthephorinae increased from 0 to 1.0, and *Trachys muricata* is now placed inside the subtribe. Also, for the stem node, q1 (which represents the quartet support value) increased from 0.42 to 0.97.

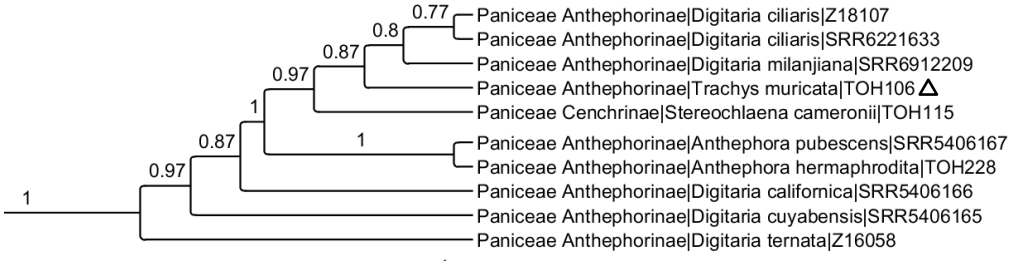


Figure 3-11: Part of 64-gene coalescent tree with PP values, showing the subtribe Anthephorinae. *Trachys muricata* is marked by a triangle.

Examples from *Polytoca digitata*, *Holcolemma canaliculatum* and *Trachys muricata* show that removal of long branches from single-gene trees and manual selection could improve the

phylogeny by purging erroneous sequences. To get a more previse position for specific species, a reduced dataset could be used, excluding the gene trees that are irrelevant.

### 3.3.3 Interpreting coalescent results – resolving extremely low support values

As shown by examples from *Polytoca digitata* and *Holcolemma canaliculatum* in 3.3.2, I hypothesized that the coalescent method (as implemented in Astral 5.7.8) tends to place a taxon at a relatively "basal" position with high support (posterior probability value, PP) if there is considerable amount of conflict among single-gene trees but no dominant topology. This can be explained by the underling methodology for coalescent analyses: it summarizes topological information from single-gene trees and report a species tree that maximize the quartet score for the whole tree instead of any specific branch. Therefore, *Polytoca digitata* was reported in such a position as shown in the original 1234-gene coalescent tree (Figure 3-7 left), even if this specific position is never observed in any single-gene trees. Without manual inspection of single-gene trees, conclusions made for such taxa are prone to errors.

To test my hypothesis regarding this misleading behavior of coalescent analyses, I performed simulations using mock data. Gene trees were made up with species of fixed relationships plus a species "X" that was *incertae sedis* (or a so-called rogue taxon). The gene trees represent all possible placements of species X and are named accordingly (see Figure 3-12 for examples). Coalescent trees were reconstructed from sets of gene trees with different compositions (data not shown).

As expected, when all types of gene trees are equally represented, the coalescent result puts X in a position sister to a combined clade of A and B taxa (Figure 3-13, set 2), with maximum support (PP=1.0), even if this topology (HAB) only takes up 1/27 of the gene trees. This is because

the number of taxa in A (five) and B (three) are higher than that in C (one), D (two) or E (one), and placing X sister to a clade of A plus B could maximize the quartet score for coalescent result. With a slightly higher portion of gene trees that place X further basal (closer to the outgroup taxon), the coalescent tree from set 1 and set 7 still place X as sister to A and B combined, but the corresponding support values are decreased (Figure 3-13, set 1 and 7).

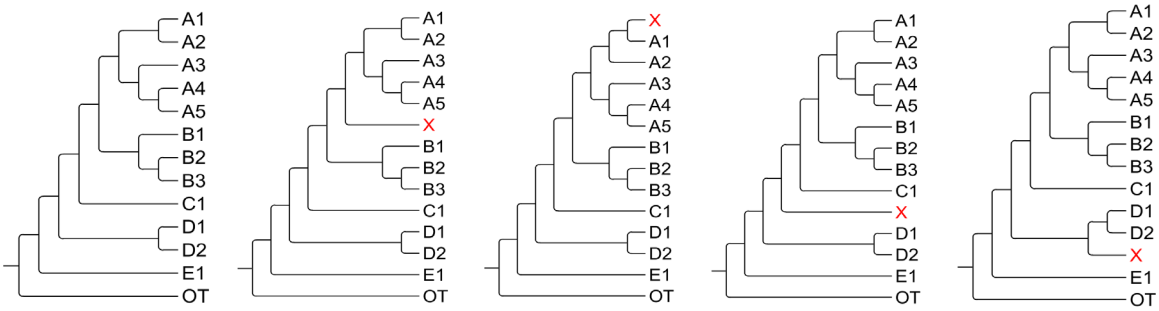


Figure 3-12: Examples of simulated gene trees with different placement of species X. The relationships of all other taxa are fixed, A through E are ingroup taxa, and OT is an outgroup. H0: X is absent in gene tree. HA: X is sister to a combined clade consisting of A1 through A5. HA1: X is sister to A1. HABC: X is sister to a combined clade consisting of all A, B and C taxa. HD: X is sister to D1 plus D2.

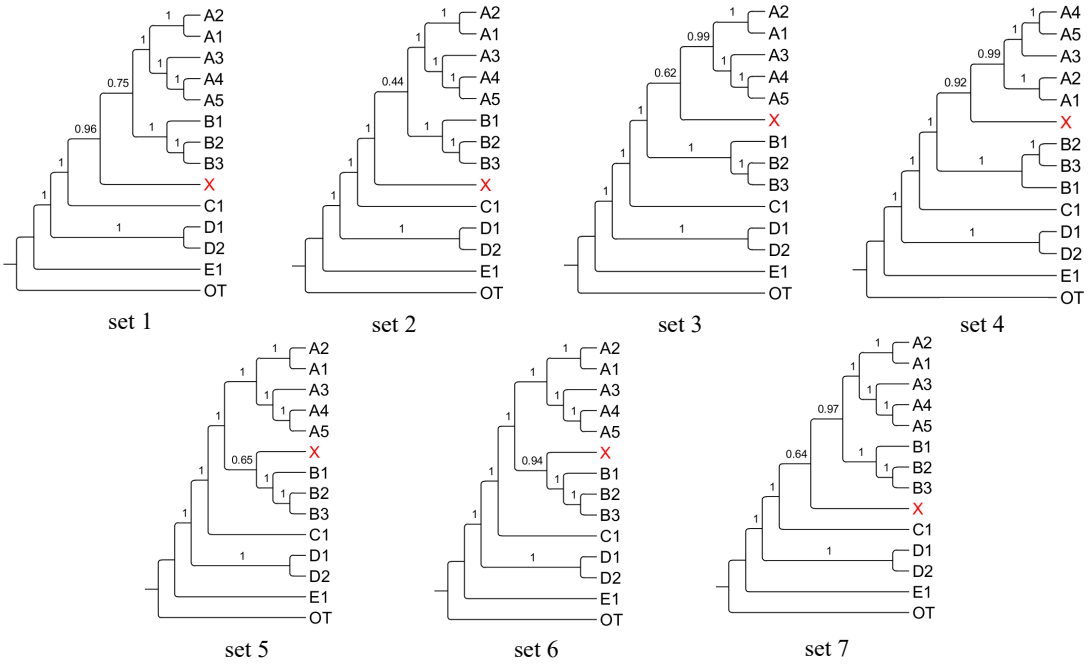


Figure 3-13: Coalescent trees from seven set of gene trees. set1: larger portion of HABCDE gene trees. set2: all types of gene trees equally represented. set3: higher portion of HA, HB type gene trees (including HA1, HB1, etc.). set4: higher portion of HA type gene trees. set5: higher portion of HB type gene trees (including HB1, HB2, etc.) set6: even higher portion of HB type gene trees. set7: higher portion of HABC, HABCD, HC, HD, etc. type gene trees.

If a higher portion of both HA and HB type gene trees (including HA1, HB1, etc.) are used as input, the coalescent result tends to place X as sister to clade A (Figure 3-13, set 3), as clade A is large than clade B, therefore conferring more "attraction". With only a higher portion of HA type gene trees used, the result shows a strong support for X as sister to clade A (Figure 3-13, set 4), with higher support than that in set 3. Similarly, increasing the portion of HB type gene trees will result in a placement of X sister to clade B, and the support value is positively correlated with the proportion of HB type gene trees (Figure 3-13, set 5 and 6).

Based on the simulations performed, the following inferences are made. For coalescent analyses, using larger portion of gene trees supporting X in a clade (i.e., A+B, A, or B) will result in X placed closer to that clade, and the support value is positively correlated with the proportion of supporting gene trees. Expanding sampling of the clade may also improve the resolution for position of species X, as this is similar to increase the proportion of gene trees supporting X in a more confined range of position. This is exemplified by comparing set 1, 2 to set 3, 4 or set 5, 6 in Figure 3-13. As the proportion of gene trees supporting X in clade A (set 3, 4) or clade B (set 5, 6) is increased, the position of X is further improved, as compared to that in set 1 or 2.

On the contrary, a more even composition of gene trees will result in X at a more basal position, and if proportionally increase different type of gene trees, X will be closer to the larger clade (Figure 3-13, set 2). Nevertheless, it is possible that the position of species X in such coalescent tree is never a dominant one (<25%) among the gene trees used, if conflicts exist as for the specific position of X among the clades. Therefore, the coalescent result may not reflect a reliable scenario even high support values are reported. In such circumstances, manual inspection of single-gene trees is necessary. Notably, the position of species X doesn't affect the relationship of other taxa in the coalescent analyses.

To conclude, if a taxon is observed at a questionable (based on prior knowledge) position in coalescent results, inspection of single-gene trees is necessary. For species from genome-

skimming data, the lower depth of sequencing data might cause a higher chance of missing rate/partial coverage for target genes, thus making the species' position poorly resolved among single-gene trees. Removal of long branches and manual inspection of gene trees could potentially purge erroneous sequences and improve the coalescent result.

Another interesting and confusing scenario is exemplified by the phylogenetic position of *Trachys muricata* (Figure 3-10, 3-11). After purging of long branches from the initial single-gene trees, the resolution for *Trachys muricata* was improved, but the support value for its associated branch on the coalescent tree was zero. Notably, the quartet support values for this node were q1=0.42, q2=0.49 and q3=0.09. Although the topology for q1 was selected, q1 is actually smaller than q2, and this is because using q1 for this node would result in a higher quartet score for the whole coalescent tree. In other words, using the 984-gene set for coalescent analyses, topological information for other taxa all needs to be taken into consideration, therefore sacrificing the resolution for this specific species, *Trachys muricata*. As a comparison, when only the 64 gene trees that contain *Trachys muricata* were used for coalescent analyses, the resolution for its position is improved (Figure 3-11), and the quartet support value q1 for this subtribe is significantly improved to 0.97, resulting in a PP value of one. Therefore, in a large-scale phylogenetic project involving hundreds of species, optimization of position for every single taxon cannot be guaranteed in one coalescent analysis, and using subsets of gene trees emphasizing specific taxa could help with obtaining higher local resolution.

### 3.3.4 An expanded gene family analysis of *ppc*

After pruning of bacterial type *ppc* genes, a total of 1119 putative plant-type *ppc* sequences from transcriptomic/genomic datasets were included for a gene family analysis. Based on previous studies (Christin et al., 2007; Christin and Besnard, 2009), Poaceae *ppc* genes were categorized into six clades: *ppc-aL1a*, *ppc-aL1b*, *ppc-aL2*, *ppc-B1*, *ppc-B2*, and *ppc-aR*. Our results confirmed the monophyly of *ppc-aL1a*, *ppc-aL1b* and *ppc-aL2,* and *ppc-aL2* is sister to *ppc-aL1a* combined with *ppc-aL1b* (topology summarized in Figure 3-14). All PACMAD and BOP subfamilies are represented in *ppc-aL2*, although there is only one sequence from Bambusoideae (*Raddia brasiliensis*) and Danthonioideae (*Schismus barbatus*), respectively, and this is probably due to lower expression level of this isoform. Unexpectedly, *ppc-aL2* was not found for Chloridoideae species *Eleusine coracana* with genomic data, although representatives from all five Chloridoideae tribes contain this gene. This is probably due to lower quality of *Eleusine coracana* genome annotation, but the probability for loss of *ppc-aL2* in this species cannot be excluded. In *ppc-aL2*, the relationship among PACMAD subfamilies is different from the established species phylogeny, although both PACMAD and BOP clades are monophyletic. For *Pharus latifolius* (Pharoideae), both RNA-seq and genomic data was used, and this species clearly has one *ppc-aL2* gene. Situation is similar in *ppc-aL1b*, but this gene seems to be missing in Danthonioideae and Pooideae. There was no genomic data available for Danthonioideae, so whether *ppc-aL1b* is lost in the genomes or not recovered due to low expression cannot be verified. For Pooideae, genomes of *Brachypodium distachyon* and *Hordeum vulgare* were included, along with transcriptomes of a few other species in this subfamily. None of these datasets contain a *ppc-aL1b* gene, so this isoform is possibly lost in Pooideae. As reported in chapter 2, four $C_4$-type *ppc* genes from Chloridoideae and Aristidoideae were found in this gene clade.

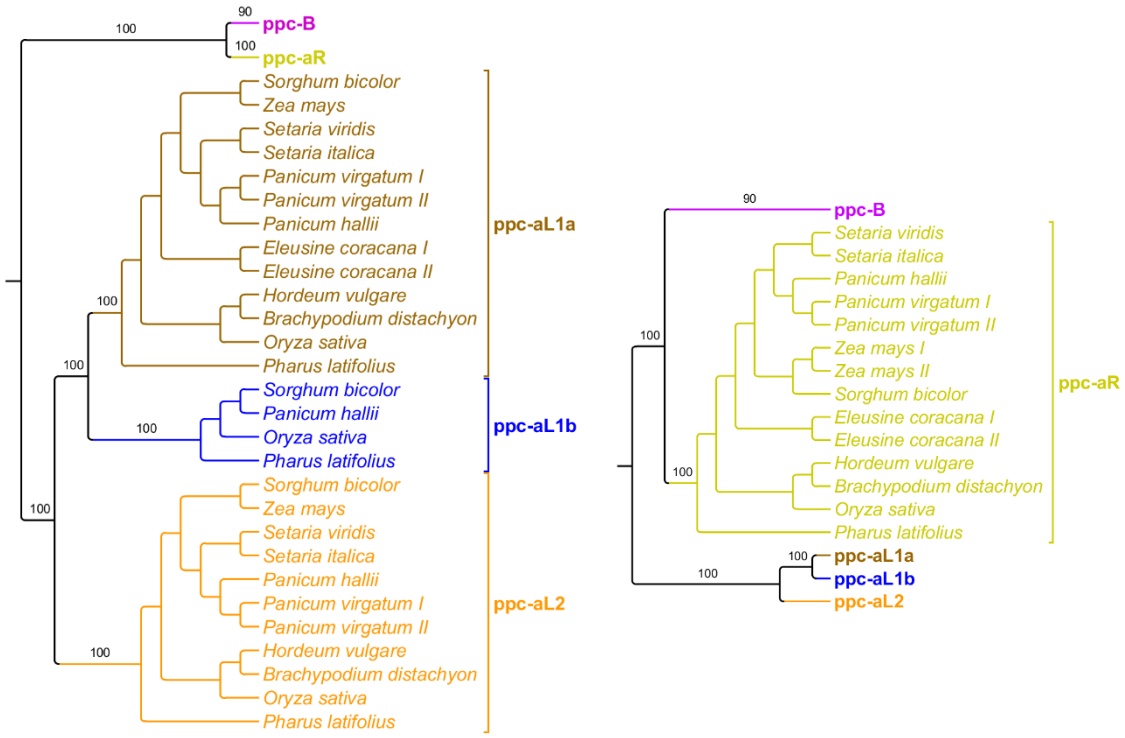


Figure 3-14: A summary of the 1119-sequence *ppc* gene family analysis. For brevity, only relationship among sequences from Poaceae genomes were shown. Bootstrap support values for each *ppc* gene clade and connecting branches are shown. For species with more than one gene in a *ppc* gene clade, sequences are labelled by roman numerals.

The *ppc-aL1a* clade is larger compared to *ppc-aL1b* or *ppc-aL2*, since this gene is found in all eleven Poaceae subfamilies sampled, and relationship among subfamilies reflected in this gene clade is largely consistent with the species phylogeny. Intriguingly, two $C_4$-type *ppc* are embedded in *ppc-aL1a*, one from *Eriachne aristidea* (Micraroideae) and the other from *Arundinella hookeri* (Panicoideae). *Arundinella hookeri* has at least one more copy of *ppc-aL1a* gene, and at least two more C4 *ppc* in *ppc-B* clade, suggesting a duplication event in this specific species or genus. Both *ppc-aL1a* and *ppc-aL1b* are found exclusively in Poaceae, suggesting that the duplication event that generated these two genes are specific to the common ancestor of grasses.

Situation is more complex in the combined clade of *ppc-B1/B2* and *ppc-aR*. Based on the gene tree with all 1119 *ppc* sequences, the *ppc-aR* clade can be clearly recognized. This is a much larger *ppc* gene clade with most species from all of PACMAD and BOP clade subfamilies. *Pharus*

*latifolius* has one *ppc-aR* gene, while the position for a cluster of sequences from *Streptochaeta* differs among preliminary trees (data not shown). Christin et al. (2007, 2009) proposed *ppc-B1* and *ppc-B2* clades based on their gene trees from exons 8 to 10 and introns. However, in their results *ppc-B1* clade is significantly smaller than *ppc-B2*, and a considerable number of genera such as *Zea*, *Setaria* and *Oryza* are missing in *ppc-B2*. Based on our gene trees, there is no strong evidence to split this clade based on gene tree topology, either from the whole dataset of 1119 sequences or from two reduced datasets containing 329 (not shown) and 30 sequences (Figure 3-17, see discussion in 3.4.2), respectively. So, we tend to be conservative and just refer to this combined "*ppc-B1"* plus "*ppc-B2*" clade as *ppc-B***.** Regardless of the delimitation, this clade contains *ppc* genes from all PACMAD and BOP subfamilies, plus *Pharus latifolius* from Pharoideae and *Streptochaeta* from Anomochlooideae, indicating this clade probably existed in the common ancestor of Poaceae. Interestingly, there is one cluster of sequences embedded in *ppc-B* clade that contains *ppc* genes from BOP subfamilies, Pharoideae and Anomochlooideae. Its position in the gene tree is weird but robust to sampling change among preliminary gene trees. Considering the composition of this cluster, it may represent an ancestral copy of the *ppc-B* gene that did not go through duplication as for PACMAD *ppc-B* genes.

The inclusion of *ppc* genes from *Pharus latifolius* genome was helpful as it clearly shows that all the five/six *ppc* clades existed at least in the common ancestor of PACMAD, BOP, plus Puelioideae and Pharoideae. The *ppc-aL1b* gene is lost in many lineages, especially in Pooideae, where none of the sampled species including both genomic and transcriptomic datasets contain this gene (Table 3-2). On the other hand, *ppc-aL2*, *ppc-aL1a* and *ppc-aR* retain a state of single-copy in most subfamilies, with additional duplications specific to some genera (e.g., *Zea*, *Panicum*). *Eleusine coracana* is worth further investigation as it does not have *ppc-aL2* but has two copies for *ppc-aL1a*, *ppc-aR* and *ppc-B*, respectively. This is consistent with the allotetraploid status of this

species, a hybrid of the diploid *Eleusine indica* (AA) with *Eleusine floccifolia* (BB) (Bisht and Mukai, 2002).

**Table 3-2: Number of plant type *ppc* genes in Poaceae species**

| subfamily | species | *ppc-aL2* | *ppc-aL1a* | *ppc-aL1b* | *ppc-aR* | *ppc-B* |
|---|---|---|---|---|---|---|
| Panicoideae | *Panicum hallii* | 1 | 1 | 1 | 1 | 1 |
| | *Panicum virgatum* | 2 | 2 | 0 | 2 | 2 |
| | *Sorghum bicolor* | 1 | 1 | 1 | 1 | 1 |
| | *Setaria italica* | 1 | 1 | 0 | 1 | 2 |
| | *Setaria viridis* | 1 | 1 | 0 | 1 | 2 |
| | *Zea mays* | 1 | 1 | 0 | 2 | 1 |
| Chloridoideae | *Eleusine coracana* | 0 | 2 | 0 | 2 | 4 |
| Pharoideae | *Pharus latifolius* | 1 | 1 | 1 | 1 | 2 |
| Pooideae | *Brachypodium distachyon* | 1 | 1 | 0 | 1 | 2 |
| | *Hordeum vulgare* | 1 | 1 | 0 | 1 | 2 |
| Oryzoideae | *Oryza sativa* | 1 | 1 | 1 | 1 | 1 |

As for Panicoideae, the four tribes Andropogoneae, Arundinelleae, Paspaleae and Paniceae each contains the all of the four genes *ppc-aL2, ppc-aL1a, ppc-aL1b* and *ppc-aR*, indicating that these genes were retained in the ancestor of Panicoideae. But *ppc-aL1b* is lost in *Panicum virgatum, Zea mays* and *Setaria,* probably due to genus-specific gene lost events. The number of *ppc* genes are doubled in *Panicum virgatum*, consistent with the ploidy level (tetraploid) of the sequenced genome (INSDC: JABWAI010000000). For $C_4$ photosynthesis, all Panicoideae species but *Arundinella hookeri* solely recruited *ppc-B* as the functional $C_4$ *ppc* gene. The origin of $C_4$ *ppc-aL1a* in *Arundinella hookeri* deserves further investigation.

### 3.3.5 Using GC content to distinguish *ppc* genes

As presented in 3.3.3, the majority of functional $C_4$ *ppc* genes are from the *ppc-B* clade, and only a small number are from *ppc-aL1a/ppc-aL1b* clade. To explain this uneven distribution, I proposed two hypotheses: (1) the *ppc-B2* genes are closer to $C_4$ function compared with *ppc-aL1a* and *ppc-aL1b* (at sequence level); (2) expression pattern (spatial and temporal) differs between these genes due to expressional regulation, such as cis-elements; in addition, expression level correlates with mutation rate, which is a proxy of the evolutionary potential, and genes of higher expression level tend to evolve faster. However, no significant difference in amino acid composition was found among the six *ppc* clades (data not shown), and more specifically there is no difference at the critical site (Ser 780 in *Zea mays* $C_4$ PEPC), if a non-$C_4$ *ppc* from *ppc-B* or *ppc-aL1a/aL1b* were to mutate into a $C_4$ version. As introduced in 3.1.2, GC content is reported to be correlated with expression level. Therefore, GC content of the five *ppc* gene clades were calculated.

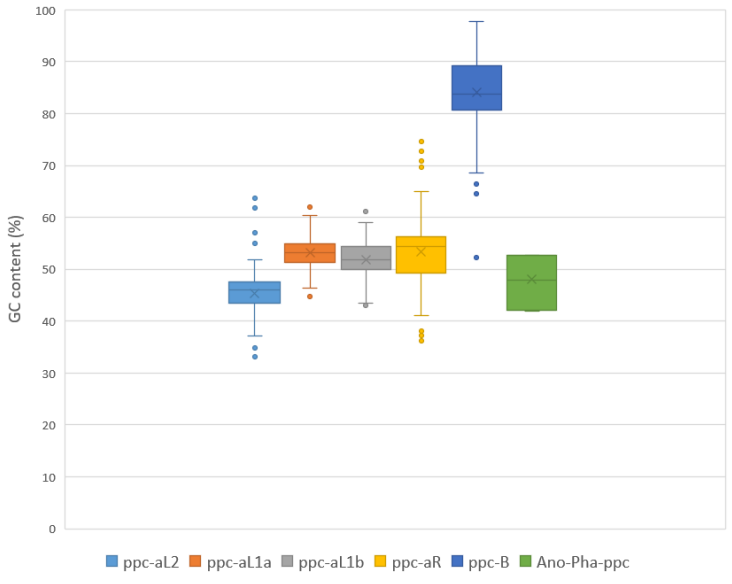


Figure 3-15: GC content at 3rd codon positions of *ppc* genes. Boxplots were generated for the five *ppc* clades, and a smaller clade containing only seven sequences from *Streptochaeta* and *Pharus*. Note that *ppc-B* has a significantly higher 3rd codon GC content compared with other *ppc* genes.

As shown in Figure 3-15, *ppc-B* has significantly higher GC content (median =83.66%) at the 3rd codon positions, compared with the other four *ppc* genes. Even for *ppc-aR* which is sister to *ppc-B*, the median 3rd codon position GC content is only 54.34%. Similarly, the more basal ppc-aL2 genes also have a lower GC content (median=45.99%). In the *ppc-B* clade, many (but not all) species have more than one copy of *ppc* genes, suggesting gene duplications in the common ancestor of some lineages. Considering previous studies (Christin and Besnard, 2009) proposed splitting this clade into *ppc-B1* and *ppc-B2*, the *ppc-B* clade likely contain two paralogous *ppc* genes.

Based on my previous observations (data not shown) of manual inspecting single-gene trees, for a specific group of putative orthologous genes (e.g., the 1,234 OGs used for phylogenetic analysis), if the GC content distribution for this groups of sequences does not follow a unimodal distribution (e.g., close to a normal distribution), then there is likely non-orthologs in the group. To test if the sequences from *ppc-B* clade are all from one single orthologous group, a distribution of GC content for all the 320 sequences were generated (Figure 3-16).

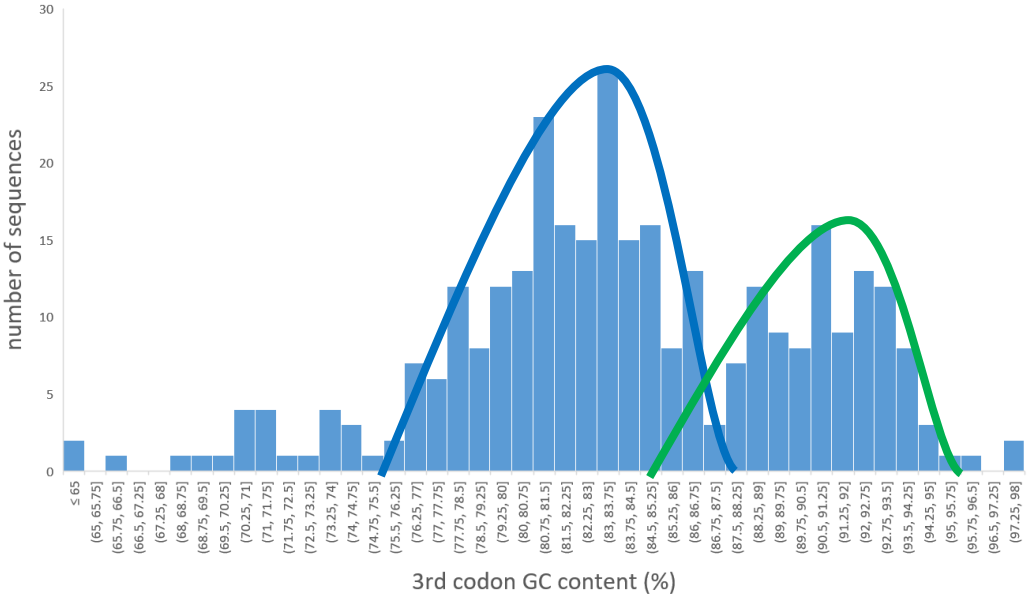


Figure 3-16: GC content distribution of *ppc-B* genes. 320 sequences were included, and the GC content ranges from 52.18~97.69%.

As shown in Figure 3-16, two peaks can be observed on the distribution. One at around 83.75%, and the second at around 91.25%. The boundary between these two peaks is not clear, probably due to overlapping tails, but a trough can be seen at around 87.5%. This is evidence to support there are two paralogous genes in this group of *ppc-B* sequences (based on my own observations among the 1,234 low-copy nuclear genes).

To further disentangle the *ppc-B* clade, a gene tree of reduced number of sequences was reconstructed, using only *ppc-B* genes from Poaceae genomes, *Streptochaeta* transcriptome and two *ppc* genes from *Ecdeiocoleae monostachya* as outgroups. This is a significant reduction for the number of sequences (from 320 in the 1119-seq *ppc* tree to 30 in this reduced set), but the result did not show a great improvement for the topology of *ppc-B* (Figure 3-17). A group of *ppc-B* from *Streptochaeta* and *Pharus* form a clade sister to the rest of ppc-B genes. This group of genes is characterized by much lower 3$^{rd}$ codon GC content (<55%), resembles more to *ppc-aR* genes in this regard. The rest of ppc-B can be classified into two categories based on 3$^{rd}$ codon GC content: lower GC (green triangle) and higher GC (blue rectangle), each falling into range ofthe two peaks in Figure 3-16. The lower GC category contain species from the basal subfamilies Anomochlooideae (*Streptochaeta*) and Pharoideae (*Pharus*), as well as species from BOP (*Hordeum* and *Brachypodium*) and PACMAD (*Eleusine, Zea, Sorghum, Setaria* and *Panicum*), while the higher GC category only contains *Oryza, Hordeum* and *Brachypodium* from BOP and *Setaria* and *Panicum* from PACMAD. The composition of two categories and gene tree topology makes the evolutionary history of *ppc-B* elusive, but a possible scenario is a duplication in the common ancestor of Poaceae followed by subsequent duplications and gene loss. On the other hand, as shown in Figure 3-17, all four *ppc-B* genes from *Elusine coracana* fall into the lower GC content category, and genes from both categories are recruited for $C_4$ function (red sequences in Figure 3-17). This indicates that the topology shown in Figure 3-17 could be impacted by selection forces, and impacted sites needs to be identified and excluded to improve the gene tree.

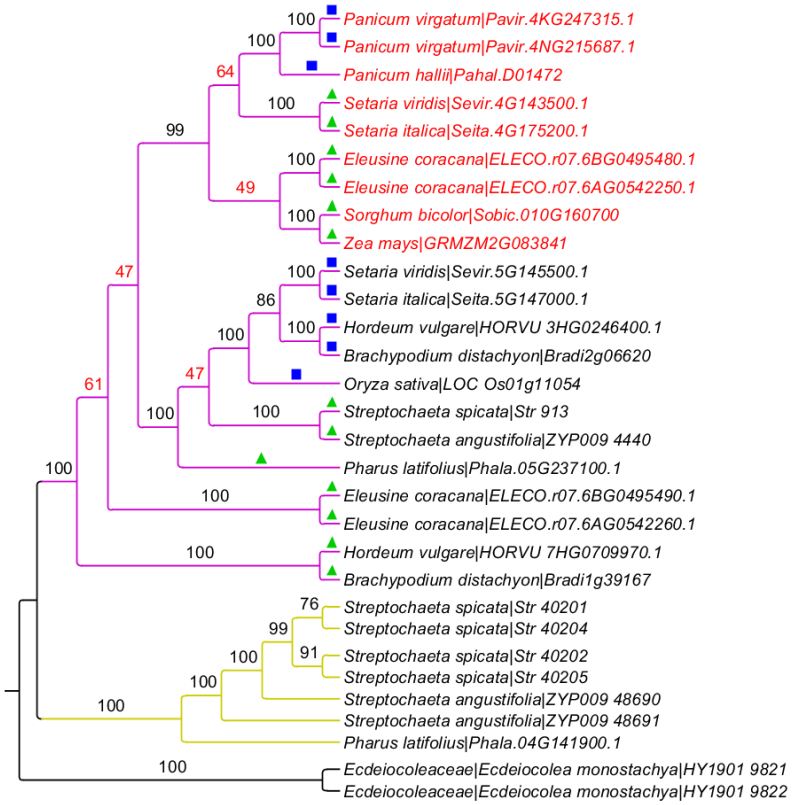


Figure 3-17: Gene tree for *ppc-B* using a reduced dataset. Two sequences from *Ecdeiocoleae monostachya* were used as outgroup, and the sequences from *Streptochaeta* were included to represent subfamily Anomochlooideae. Functional $C_4$ *ppc* genes are in red fonts, and sequences are labelled according to GC content level: yellow branch, lower than 55%; green triangle, falls into the first group in Figure 3-17; blue rectangle, falls into the second group in Figure 3-17.

Nevertheless, using GC content as an indicator, evidence is gathered to support that there are at least two paralogous genes in the *ppc-B* clade. The overall higher 3$^{rd}$ codon GC content for *ppc-B* genes is probably a derived trait in both copies for core Poaceae and in one of the two copies of *ppc-B* for Anomochlooideae and Pharoideae. A higher GC content could potentially render *ppc-B* genes higher expression level compared to other *ppc* genes, making it more suitable for C4 function. To test whether *ppc-B* genes possess higher expression potential, codon adaptation index was calculated for *ppc* genes in representative species. There are synonymous codons for amino acids, and for each set of synonymous codons some (usually one) of them are used more frequently to code for that amino acid. Codon adaptation index (CAI) values range from 0 to 1, with higher

values indicating a bigger proportion of the most abundant synonymous codons (Sharp and Li, 1987). CAI value of a specific coding sequence is calculated by comparing the codon usage frequency of itself to the overall codon usage frequency of highly expressed genes of the species. CAI is a measurement of the relative adaptiveness of the codon usage of a gene and can be used as a factor to predict expression level.

As shown in Table 3-3, in the five species included, *ppc-B* genes always have the highest CAI values, indicating they are potentially more highly expressed compared to other *ppc* genes. Thus, the higher 3$^{rd}$ codon GC content in *ppc-B* is mainly a result from synonymous codons with higher GC content. The choice of codons can influence local translation kinetics during protein synthesis. Hia et al. (2019) showed that in humans RNA binding proteins regulate mRNA half-life, depending on GC content and codon usage. RNAs with shorter half-lives were associated with AT3 codons, while those with longer half-lives were associated with GC3 codons. Newman et al. (2016) reported that codon bias and GC content contributes to the different expression levels of TLR7 and TLR9 (genes related to immunity in human), and that the major factor causing the difference is transcription rate. They proposed that suboptimal codon bias, which correlates with lower guanine-cytosine (GC) content, limits transcription of certain genes. Therefore, the higher 3$^{rd}$ codon GC content in *ppc-B* is probably an advantage, since $C_4$ photosynthesis requires sufficient among of PEPC during carbon fixation. The next question to answer is how *ppc-B* gained a higher GC content, from selection or alongside with its chromosome background.

**Table 3-3: CAI values of *ppc* coding sequences in Poaceae species**

| | *Zea mays* | *Oryza sativa* | *Eragrostis walteri* | *Bouteloua dimorpha* | *Briza maxima* |
|---|---|---|---|---|---|
| ***ppc-aL2*** | 0.731 | 0.724 | 0.860 | 0.820 | N/A |
| ***ppc-aL1a*** | 0.750 | 0.805 | 0.879 | 0.854, 0.856 | 0.833 |
| ***ppc-aL1b*** | N/A | 0.801 | 0.885 | N/A | N/A |
| ***ppc-aR*** | 0.741, 0.742 | 0.774 | 0.878 | 0.847 | 0.835 |
| ***ppc-B*** | **0.870 (C4 *ppc*)** | **0.918** | **0.890, 0.893** | **0.922 (C4 *ppc*)** | **0.892, 0.914** |

*N/A: gene not found in this sample. Codon usage databases were either retrieved from https://www.kazusa.or.jp/codon/ or calculated using non-redundant CDS datasets from our project.

The fact that $C_4$ *ppc* genes exist in *ppc-aL1a* and *ppc-aL1b* supports multiple origins of $C_4$ photosynthesis, and also reminds us of a complex history of this gene family. Both *Arundinella hookeri* and *Eriachne aristideae* has one $C_4$ *ppc-aL1a*, and they each has additional $C_4$ *ppc* from *ppc-B* clade. On the other hand, *Stipagrostis pennata* and three Chloridoideae species, all of which has $C_3$ *ppc-aL1b*, do not contain additional $C_4$ ppc from *ppc-B*. This seemingly redundance could be an alternative resource for expression of $C_4$ *ppc* in different tissues or might reflect remnants of a transition from using *ppc-aL1a/ppc-aL1b* to *ppc-B* as for $C_4$ function. Further investigation is needed to answer these intriguing questions.

## 3.4 Discussion

### 3.4.1 Challenges and benefits of incorporating genome skimming data into large-scale nuclear phylogeny

The term genome skimming technically refers to sequencing a genome with shallower depth and lower coverage. Herbarium samples can be used for this approach, making it easier to include species that are hard to require, greatly reducing the cost of obtaining fresh materials for transcriptome sequencing. However, there are also limitations for genome skimming, and the first and most relevant one as for phylogenetic studies is the lower coverage of target genes. Due to shallower depth and uneven coverage across the genome, some target genes may not be recovered, as these target genes were not chosen specifically to accommodate genome skimming data. In this respect, target enrichment method may outperform genome skimming, because for the former one the target genes were chosen and tested across different genomes to ensure their producibility. Moreover, for those target genes that can be recovered, the effective length (when align the coding regions with full-length CDS from transcriptome and genomic datasets) tend to be shorter. In my project, gene trees reconstructed from a considerable portion (~50%) of genes from genome skimming datasets are prone to erroneous results, although automatic methods and manual inspection (see 3.2.3) could purge erroneous sequences.

The second challenge for using genome skimming data is to predict the structure of target genes, as for most of the species genome annotation information is scarce. In my project HaMStR is used to obtain coding sequences from *de novo* assembled genomic contigs, which are supposed to contain additional elements other than exons, such as introns and intergenic regions. The algorithm implemented by HaMStR will try to generate possible translations from the genomic contigs, compare the results with the template sequences, and output one or multiple coding sequences predicted to be homologous with the templates. This process usually results in a shorter

coding sequence compared with the templates, since the contigs are not guaranteed to cover the full length of CDS. However, when combined with sequences from transcriptomic data and aligned, the retaining coding sequences from genome skimming datasets generally looks reliable, although missing regions is common.

The third challenge is to identify false positive sequences from the putative orthologous groups. HaMStR (or other automatic methods to predict homologs) may report sequences that are chimera of different paralogs (due to assembly errors) or sequences with only a few matched regions of templates scattered and are not true homologs. These sequences cannot be easily identified from the alignment by automatic methods, and manual inspection is needed. In my project, single-gene trees and the corresponding alignments went through auto-detection of long branches as well as manual inspection to remove sequences that are potentially paralogs or of lower quality. This has proven effective to improve the quality of single-gene trees and thus the reliability of coalescent results.

To summarize, the challenges for incorporating genome skimming data into phylogenetic analyses come from three aspects, and eventually all lead to fewer and shorter target genes. If using together with sequences of higher quality from transcriptomic or genomic datasets, these defects can be partially counteracted, but effort is necessary to inspect the alignment and corresponding gene trees. Expanding the pool of target genes is also a potential solution to compensate for lower data recovery of genome skimming data, although precautions must be taken to select for genes that are suitable for phylogenetic analyses (i.e., low-copy, putative orthologous). For species of critical phylogenetic positions (e.g., a single species representing a tribe or subtribe), increasing the sequencing depth would be beneficial, as this will improve the *de novo* assembly and hence increase the chance of recovering target genes.

### 3.4.2 The number of $C_4$ origins in Panicoideae

As a popular topic, the origin of $C_4$ photosynthesis in Poaceae has been widely discussed in previous studies. Based on a plastid gene phylogeny, GPWG II (2012) inferred 22~24 times of $C_4$ origins in Poaceae, and 17~19 of them are in Panicoideae. While their results were congruent with my nuclear-gene phylogeny for the relationship among tribes of Panicoideae s.s. which contains the majority of $C_4$ Panicoid grasses, Tristachyideae was reported as sister to Centotheceae, Cyperochloeae plus Thysanolaeneae (see Figure 3-1 B), making $C_4$ in Tristachyideae seems like an independent origin. In my results, however, Tristachyideae is consistently sister to Panicoideae s.s., therefore a more reasonable interpretation is that $C_4$ only originated in this combined clade.

Tribes Paniceae and Paspaleae are mixture of $C_3$ and $C_4$ species, where both types are intertwined wtihin subtribes. Ancestral state reconstruction based on statistical analysis tend to report multiple origins and reversals based on photosynthetic types mapped on the phylogeny, but this kind of results is merely a simplified deduction, because the conversion from one type to the other is complex and should not be modelled by a simple process. Given that $C_4$ tribe Tristachyideae is sister to all the remaining $C_4$ clade, a more reasonable hypothesis would be that preconditions for $C_4$ were developed in the common ancestor of Tristachyideae plus Panicoideae s.s., and some lineages continued on their way to $C_4$ while others retained $C_3$ (or reversed back to $C_3$) as this whole clade diverged. Therefore, previous studies possibly overestimated the number of $C_4$ origins in Panicoideae, and the position of Tristachyideae as reported in my project is critical for this question.

Due to the lower recovery rate of genes, genome skimming datasets were not included in the *ppc* gene family analysis, although many of them represent C3 species in Paniceae and Paspaleae. Nevertheless, efforts were made to obtain ppc genes from *de novo* assembled genomic contigs, and one interesting observation is a putative $C_4$ type *ppc* from *Arundoclaytonia dissimilis* (data not shown), a species close to Zeugiteae and Chasmanthieae, both reported to be wholly $C_3$.

This corresponding amino acid sequence has a Serine at #780 (*Zea mays* PEPC numbering), a characteristic of functionally $C_4$ PEPC; on the other hand, no $C_4$ type *ppc* was found in other species in Zeugiteae or Chasmanthieae. Further investigation from transcriptome sequencing or PCR-based sequencing would be needed to check if this is a functional activate gene or a pseudogene.

# Appendix A

## Fossils used for calibration in molecular clock analysis

| # | Clade | Min/Max | Age (million years) | Fossil type | References |
|---|---|---|---|---|---|
| 1 | CG_commelinids | Max | 118 | Secondary calibration | (Hertweck et al., 2015) |
| 2 | CG_Poaceae | Min | 101 | Silicified epidermal pieces and phytoliths | (Wu et al., 2018) |
| 3 | SG_Zingiberales | Min | 77 | Seeds | (Iles et al., 2015) |
| 4 | SG_*Chusquea* | Min | 35 | Phytolith | (Strömberg, 2005; Prasad et al., 2011) |
| 5 | CG_Pooideae | Min | 40 | Phytolith | (Zucol et al., 2010; Prasad et al., 2011; Iles et al., 2015) |
| 6 | SG_Stipeae | Min | 34 | Fruits | (Manchester, 2001; Iles et al., 2015) |
| 7 | SG_PACMAD | Min | 40 | Phytolith | (Zucol et al., 2010) |
| 8 | CG_Chloridoideae | Min | 19 | Phytolith | (Strömberg, 2005) |
| 9 | CG_C4_Panicoideae | Min | 12 | Macrofossil | (Nambudiri et al., 1978; Whistler et al., 2009; Prasad et al., 2011) |
| 10 | CG_Oryzoideae | Min | 66 | Epidermis and phytolith | (Prasad et al., 2011; Iles et al., 2015) |
| 11 | SG_*Leersia* | Min | 30.44 | Inflorescence | (Walther and Kvaček, 2007; Iles et al., 2015) |
| 12 | CG_(*Oryza*+*Zea*) | Min | 55 | Inflorescence, spikelet and pollen | (Crepet and Feldman, 1991) |

| | | | | | |
|---|---|---|---|---|---|
| 13 | SG_*Neyraudia* | Min | 19 | Phytolith | (Dugas and Retallack, 1993; Strömberg, 2005) |

*CG=crown group; SG=stem group

# Appendix B

## Molecular clock estimates of mean ages and 95% confidence intervals at major nodes in Poaceae phylogeny

| clade | low | mean | high | note |
|---|---|---|---|---|
| Crown group Poaceae | 101 | 101 | 101 | fixed |
| Crown group *Streptochaeta* (Anomochlooideae) | 6.67 | 7.03 | 7.35 | |
| Crown group *Guaduella* (Puelioideae) | 46.01 | 48.19 | 49.58 | |
| Crown group *Puelia* (Puelioideae) | 42.56 | 44.48 | 47.12 | |
| Crown group (BOP + PACMAD) | 81.17 | 81.43 | 81.78 | |
| Crown group BOP | 77.69 | 77.94 | 78.29 | |
| Crown group subfamily Oryzoideae | 66 | 66 | 66 | fixed |
| Crown group subfamily Bambusoideae | 66.46 | 66.89 | 67.32 | |
| Crown group subfamily Pooideae | 66.54 | 67.03 | 67.5 | |
| Crown group PACMAD | 65.98 | 66.33 | 66.83 | |
| Crown group subfamily Aristidoideae | 46.93 | 47.46 | 48.28 | |
| Crown group subfamily Micrairoideae | 32.71 | 33.07 | 33.68 | |
| Crown group subfamily Panicoideae | 53.72 | 54.22 | 54.78 | |
| Crown group subfamily Arundinoideae | 56.84 | 57.39 | 57.97 | |
| Crown group subfamily Danthonioideae | 35.15 | 35.8 | 36.51 | |
| Crown group subfamily Chloridoideae | 57.31 | 57.76 | 58.22 | |

# References

**Aggarwal, R. K., Brar, D. S., Nandi, S., Huang, N., and Khush, G. S.** (1999). Phylogenetic relationships among *Oryza* species revealed by AFLP markers. Theor. Appl. Genet. **98**:1320–1328.

**Aliscioni, S. S., Giussani, L. M., Zuloaga, F. O., and Kellogg, E. A.** (2003). A molecular phylogeny of *Panicum* (Poaceae: Paniceae): Tests of monophyly and phylogenetic placement within the Panicoideae. Am. J. Bot. **90**:796–821.

**Aliscioni, S., Bell, H. L., Besnard, G., Christin, P. A., Columbus, J. T., Duvall, M. R., Edwards, E. J., Giussani, L., Hasenstab-Lehman, K., Hilu, K. W., et al.** (2012). New grass phylogeny resolves deep evolutionary relationships and discovers C4 origins. New Phytol. **193**:304–312.

**Andrews, S., Krueger, F., Seconds-Pichon, A., Biggins, F., and Wingett, S.** (2015). FastQC. A quality control tool for high throughput sequence data. Babraham Inst. **1**:1.

**Barker, N. P., Linder, H. P., Morton, C. M., and Lyle, M.** (2003). The paraphyly of *Cortaderia* (Danthonioideae; Poaceae): Evidence from morphology and chloroplast and nuclear DNA sequence data. Ann. Missouri Bot. Gard. **90**:1–24.

**Barrett, C. F., Baker, W. J., Comer, J. R., Conran, J. G., Lahmeyer, S. C., Leebens-Mack, J. H., Li, J., Lim, G. S., Mayfield-Jones, D. R., Perez, L., et al.** (2016). Plastid genomes reveal support for deep phylogenetic relationships and extensive rate variation among palms and other commelinid monocots. New Phytol. **209**:855–870.

**Bayzid, M. S., Mirarab, S., Boussau, B., and Warnow, T.** (2015). Weighted statistical binning: Enabling statistically consistent genome-scale phylogenetic analyses. PLoS One **10**.

**Benz, B. F.** (2001). Archaeological evidence of teosinte domestication from Guilá Naquitz, Oaxaca. Proc. Natl. Acad. Sci. U. S. A. **98**:2104–2106.

**Besnard, G., Pinçon, G., D'Hont, A., Hoarau, J. Y., Cadet, F., and Offmann, B.** (2003). Characterisation of the phosphoenolpyruvate carboxylase gene family in sugarcane (*Saccharum* spp.). Theor. Appl. Genet. **107**:470–478.

**Bianconi, M. E., Hackel, J., Vorontsova, M. S., Alberti, A., Arthan, W., Burke, S. V, Duvall, M. R., Kellogg, E. A., Lavergne, S., McKain, M. R., et al.** (2020). Continued adaptation of C4 photosynthesis after an initial burst of changes in the Andropogoneae grasses. Syst. Biol. **69**:445–461.

**Bisht, M. S., and Mukai, Y.** (2002). Genome organization and polyploid evolution in the genus *Eleusine* (Poaceae). Plant Syst. Evol. **233**:243–258.

**Bläsing, O. E., Westhoff, P., and Svensson, P.** (2000). Evolution of C4 phosphoenolpyruvate carboxylase in *Flaveria*, a conserved serine residue in the carboxyl-terminal part of the enzyme is a major determinant for C4-specific characteristics. J. Biol. Chem. **275**:27917–27923.

**Bolger, a. M., Lohse, M., and Usadel, B.** (2014). Trimmomatic: A flexible read trimming tool for Illumina NGS data. Bioinformatics **30**:2114–2120.

**Bouchenak-Khelladi, Y., Salamin, N., Savolainen, V., Forest, F., Bank, M. van der, Chase, M. W., and Hodkinson, T. R.** (2008). Large multi-gene phylogenetic trees of the grasses (Poaceae): Progress towards complete tribal and generic level sampling. Mol. Phylogenet. Evol. **47**:488–505.

**Bouchenak-Khelladi, Y., Muasya, A. M., and Linder, H. P.** (2014). A revised evolutionary history of Poales: Origins and diversification. Bot. J. Linn. Soc. **175**:4–16.

**Bowers, J. E., Tang, H., Burke, J. M., and Paterson, A. H.** (2022). GC content of plant genes is linked to past gene duplications. PLoS One **17**.

**Briggs, B. G., Marchant, A. D., and Perkins, A. J.** (2014). Phylogeny of the restiid clade (Poales) and implications for the classification of Anarthriaceae, Centrolepidaceae and Australian Restionaceae. Taxon **63**:24–46.

**Butte, A. J., Dzau, V. J., and Glueck, S. B.** (2001). Further defining housekeeping, or "maintenance," genes Focus on "A compendium of gene expression in normal human tissues". Physiol. Genomics **7**:95–96.

**Capella-Gutiérrez, S., Silla-Martínez, J. M., and Gabaldón, T.** (2009). trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics **25**:1972–1973.

**Cerros-Tlatilpa, R., and Columbus, J. T.** (2009). C3 photosynthesis in *Aristida longifolia*: Implication for photosynthetic diversification in Aristidoideae (Poaceae). Am. J. Bot. **96**:1379–1387.

**Chemisquy, M. A., Giussani, L. M., Scataglini, M. A., Kellogg, E. A., and Morrone, O.** (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.

**Cheng, F., Wu, J., Cai, X., Liang, J., Freeling, M., and Wang, X.** (2018). Gene retention, fractionation and subgenome differences in polyploid plants. Nat. Plants **4**:258–268.

**Chikhi, R., and Medvedev, P.** (2014). Informed and automated k-mer size selection for genome assembly. Bioinformatics **30**:31–37.

**Choi, J. Y., Zaidem, M., Gutaker, R., Dorph, K., Singh, R. K., and Purugganan, M. D.** (2019). The complex geography of domestication of the African rice *Oryza glaberrima*. PLoS Genet. **15**.

**Christenhusz, M. J. M., and Byng, J. W.** (2016). The number of known plants species in the world and its annual increase. Phytotaxa **261**:201–217.

**Christin, P. A., and Besnard, G.** (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., Salamin, N., Savolainen, V., Duvall, M. R., and Besnard, G.** (2007a). C4 photosynthesis evolved in grasses via parallel adaptive genetic changes. Curr. Biol. **17**:1241–1247.

**Christin, P. A., Salamin, N., Savolainen, V., and Besnard, G.** (2007b). A phylogenetic study of the phosphoenolpyruvate carboxylase multigene family in Poaceae: Understanding the molecular changes linked to C4 photosynthesis evolution. Kew Bull. **62**:455–462.

**Christin, P. A., Freckleton, R. P., and Osborne, C. P.** (2010). Can phylogenetics identify C4 origins and reversals? Trends Ecol. Evol. **25**:403–409.

**Christin, P. A., Edwards, E. J., Besnard, G., Boxall, S. F., Gregory, R., Kellogg, E. A., Hartwell, J., and Osborne, C. P.** (2012). Adaptive evolution of C4 photosynthesis through recurrent lateral gene transfer. Curr. Biol. **22**:445–449.

**Christin, P. A., Spriggs, E., Osborne, C. P., Strömberg, C. A. E., Salamin, N., and Edwards, E. J.** (2014). Molecular dating, evolutionary rates, and the age of the grasses. Syst. Biol. **63**:153–165.

**Christin, P. A., Arakaki, M., Osborne, C. P., and Edwards, E. J.** (2015). Genetic enablers underlying the clustered evolutionary origins of C4 photosynthesis in angiosperms. Mol. Biol. Evol. **32**:846–858.

**Clark, L. G., and Judziewicz, E. J.** (1996). The grass subfamilies Anomochlooideae and Pharoideae (Poaceae). Taxon **45**:641–645.

**Clark, L. G., Zhang, W., and Wendel, J. F.** (1995). A phylogeny of the grass family (Poaceae) based on *ndhF* sequence data. Syst. Bot. **20**:436–460.

**Clark, L. G., Kobayashi, M., Mathews, S., Spangler, R. E., and Kellogg, E. A.** (2000). The Puelioideae, a new subfamily of Poaceae. Syst. Bot. **25**:181–187.

**Columbus, T.** (1999). An expanded circumscription of Bouteloua (Gramineae: Choridoideae): new combinations and names. Aliso **18**:61–65.

**Columbus, T., Cerros-Tlatilpa, R., Kinney, M., Siqueiros-Delgado, M., Bell, H., Griffith, P., and Refulio-Rodriguez, N.** (2007). Phylogenetics of Chloridoideae (Gramineae): a Preliminary Study Based on Nuclear Ribosomal Internal Transcribed Spacer and Chloroplast trnL–F Sequences. Aliso **23**:565–579.

**Cotton, J. L., Wysocki, W. P., Clark, L. G., Kelchner, S. A., Pires, J. C., Edger, P. P., Mayfield-Jones, D., and Duvall, M. R.** (2015). Resolving deep relationships of PACMAD grasses: A phylogenomic approach. BMC Plant Biol. **15**.

**Crayn, D. M., Winter, K., and Smith, J. A. C.** (2004). Multiple origins of crassulacean acid metabolism and the epiphytic habit in the Neotropical family Bromeliaceae. Proc. Natl. Acad. Sci. U. S. A. **101**:3703–3708.

**Crepet, W. L., and Feldman, G. D.** (1991). The earliest remains of grasses in the fossil record. Am. J. Bot. **78**:1010–1014.

**Davidse, G., and Ellis, R. P.** (1987). *Arundoclaytonia*, A New Genus of the Steyermakochloeae (Poaceae: Arundinoideae) From Brazil. Ann. Missouri Bot. Gard. **74**:479.

**Davis, J. I., and Soreng, R. J.** (1993). Phylogenetic structure in the grass family (Poaceae) as inferred from chloroplast DNA restriction site variation. Am. J. Bot. **80**:1444–1454.

**Dong, L., Patil, S., Condon, S. A., Haas, E. J., and Chollet, R.** (1999). The conserved C-terminal tetrapeptide of sorghum C4 phosphoenolpyruvate carboxylase is indispensable for maximal catalytic activity, but not for homotetramer formation. Arch. Biochem. Biophys. **371**:124–128.

**Dugas, D. P., and Retallack, G. J.** (1993). Middle Miocene fossil grasses from Fort Ternan, Kenya. J. Paleontol. **67**:113–128.

**Dujardin, M.** (1978). Chromosome numbers of some tropical African grasses from western Zaïre. Can. J. Bot. **56**:2138–2152.

**Ebersberger, I., Strauss, S., and Von Haeseler, A.** (2009). HaMStR: Profile hidden markov model based search for orthologs in ESTs. BMC Evol. Biol. **9**:157.

**Edwards, E. J., and Smith, S. A.** (2010). Phylogenetic analyses reveal the shady history of C4 grasses. Proc. Natl. Acad. Sci. U. S. A. **107**:2532–2537.

**Edwards, E. J., and Still, C. J.** (2008). Climate, phylogeny and the ecological distribution of C4 grasses. Ecol. Lett. **11**:266–276.

**Estep, M. C., McKain, M. R., Diaz, D. V., Zhong, J., Hodge, J. G., Hodkinson, T. R., Layton, D. J., Malcomber, S. T., Pasquet, R., and Kellogg, E. A.** (2014). Allopolyploidy, diversification, and the Miocene grassland expansion. Proc. Natl. Acad. Sci. U. S. A. **111**:15149–15154.

**Feldman, M., and Kislev, M. E.** (2007). Domestication of emmer wheat and evolution of free-threshing tetraploid wheat. In *Israel Journal of Plant Sciences*, pp. 207–221.

**Fisher, A. E., Hasenstab, K. M., Bell, H. L., Blaine, E., Ingram, A. L., and Columbus, J. T.** (2016). Evolutionary history of chloridoid grasses estimated from 122 nuclear loci. Mol. Phylogenet. Evol. **105**:1–14.

**Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W.** (2012). CD-HIT: Accelerated for clustering the next-generation sequencing data. Bioinformatics **28**:3150–3152.

**Galtier, N., Piganeau, G., Mouchiroud, D., and Duret, L.** (2001). GC-content evolution in mammalian genomes: The biased gene conversion hypothesis. Genetics **159**:907–911.

**Ge, S., Sang, T., Lu, B. R., and Hong, D. Y.** (1999). Phylogeny of rice genomes with emphasis on origins of allotetraploid species. Proc. Natl. Acad. Sci. U. S. A. **96**:14400–14405.

**Gehrig, H., Heute, V., and Kluge, M.** (2001). New partial sequences of phosphoenolpyruvate carboxylase as molecular phylogenetic markers. Mol. Phylogenet. Evol. **20**:262–274.

**Giussani, L. M., Cota-Sánchez, J. H., Zuloaga, F. O., and Kellogg, E. A.** (2001). A molecular phylogeny of the grass subfamily Panicoideae (Poaceae) shows multiple origins of C4 photosynthesis. Am. J. Bot. **88**:1993–2012.

**Glémin, S., Scornavacca, C., Dainat, J., Burgarella, C., Viader, V., Ardisson, M., Sarah, G., Santoni, S., David, J., and Ranwez, V.** (2019). Pervasive hybridizations in the history of wheat relatives. Sci. Adv. **5**.

**Goh, W. L., Chandran, S., Franklin, D. C., Isagi, Y., Koshy, K. C., Sungkaew, S., Yang, H. Q., Xia, N. H., and Wong, K. M.** (2013). Multi-gene region phylogenetic analyses suggest reticulate evolution and a clade of Australian origin among paleotropical woody bamboos (Poaceae: Bambusoideae: Bambuseae). Plant Syst. Evol. **299**:239–257.

**Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al.** (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. **29**:644–652.

**Griffiths, H.** (1989). Carbon Dioxide Concentrating Mechanisms and the Evolution of CAM in Vascular Epiphytes.pp. 42–86.

**Guo, Z. H., Ma, P. F., Yang, G. Q., Hu, J. Y., Liu, Y. L., Xia, E. H., Zhong, M. C., Zhao, L., Sun, G. L., Xu, Y. X., et al.** (2019). Genome sequences provide insights into the reticulate origin and unique traits of woody bamboos. Mol. Plant **12**:1353–1365.

**Guo, C., Ma, P. F., Yang, G. Q., Ye, X. Y., Guo, Y., Liu, J. X., Liu, Y. L., Eaton, D. A. R., Guo, Z. H., and Li, D. Z.** (2021). Parallel ddRAD and genome skimming analyses reveal a radiative and reticulate evolutionary history of the temperate bamboos. Syst. Biol. **70**:756–773.

**Hamoud, M. A., Haroun, S. A., MacLeod, R. D., and Richards, A. J.** (1994). Cytological relationships of selected species of *Panicum* L. Biol. Plant. **36**:37–45.

**Hershberg, R., and Petrov, D. A.** (2010). Evidence that mutation is universally biased towards AT in bacteria. PLoS Genet. **6**.

**Hertweck, K. L., Kinney, M. S., Stuart, S. A., Maurin, O., Mathews, S., Chase, M. W., Gandolfo, M. A., and Pires, J. C.** (2015). Phylogenetics, divergence times and diversification from three genomic partitions in monocots. Bot. J. Linn. Soc. **178**:375–393.

**Hia, F., Yang, S. F., Shichino, Y., Yoshinaga, M., Murakawa, Y., Vandenbon, A., Fukao, A., Fujiwara, T., Landthaler, M., Natsume, T., et al.** (2019). Codon bias confers stability to human mRNA s . EMBO Rep. **20**.

**Huang, C. H., Sun, R., Hu, Y., Zeng, L., Zhang, N., Cai, L., Zhang, Q., Koch, M. A., Al-Shehbaz, I., Edger, P. P., et al.** (2016a). Resolution of Brassicaceae phylogeny using nuclear genes uncovers nested radiations and supports convergent morphological evolution. Mol. Biol. Evol. **33**:394–412.

**Huang, C. H., Zhang, C., Liu, M., Hu, Y., Gao, T., Qi, J., and Ma, H.** (2016b). Multiple polyploidization events across Asteraceae with two nested events in the early history revealed by nuclear phylogenomics. Mol. Biol. Evol. **33**:2820–2835.

**Huang, W., Zhang, L., Columbus, J. T., Hu, Y., Zhao, Y., Tang, L., Guo, Z., Chen, W., McKain, M., Bartlett, M., et al.** (2022). A well-supported nuclear phylogeny of Poaceae and implications for the evolution of C4 photosynthesis. Mol. Plant **15**:755–777.

**Hurst, L. D., and Merchant, A. R.** (2001). High guanine-cytosine content is not an adaptation to high temperature: A comparative analysis amongst prokaryotes. Proc. R. Soc. B Biol. Sci. **268**:493–497.

**Huson, D. H., and Scornavacca, C.** (2012). Dendroscope 3: An interactive tool for rooted phylogenetic trees and networks. Syst. Biol. **61**:1061–1067.

**Iles, W. J. D., Smith, S. Y., Gandolfo, M. A., and Graham, S. W.** (2015). Monocot fossils suitable for molecular dating analyses. Bot. J. Linn. Soc. **178**:346–374.

**Ingram, A. L., Christin, P. A., and Osborne, C. P.** (2011). Molecular phylogenies disprove a hypothesized C4 reversion in *Eragrostis walteri* (Poaceae). Ann. Bot. **107**:321–325.

**Jones, S. S., Burke, S. V., and Duvall, M. R.** (2014). Phylogenomics, molecular evolution, and estimated ages of lineages from the deep phylogeny of Poaceae. Plant Syst. Evol. **300**:1421–1436.

**Kai, Y., Matsumura, H., and Izui, K.** (2003). Phosphoenolpyruvate carboxylase: Three-dimensional structure and molecular mechanisms. Arch. Biochem. Biophys. **414**:170–179.

**Katoh, K., Asimenos, G., and Toh, H.** (2009). Multiple alignment of DNA sequences with MAFFT. Methods Mol. Biol. **537**:39–64.

**Kellogg, E. A.** (2009). The evolutionary history of Ehrhartoideae, Oryzeae, and *Oryza*. Rice **2**:1–14.

**Kellogg, E. A.** (2013). C4 photosynthesis. Curr. Biol. **23**:R594–R599.

**Kellogg, E. A.** (2015). Flowering plants. Monocots: Poaceae. In *Flowering Plants. Monocots: Poaceae* (ed. Kubitzki, K.), pp. 1–416.

**Kiktev, D. A., Sheng, Z., Lobachev, K. S., and Petes, T. D.** (2018). GC content elevates mutation and recombination rates in the yeast Saccharomyces cerevisiae. Proc. Natl. Acad. Sci. U. S. A. **115**:E7109–E7118.

**Kim, G., LeBlanc, M. L., Wafula, E. K., DePamphilis, C. W., and Westwood, J. H.** (2014). Genomic-scale exchange of mRNA between a parasitic plant and its hosts. Science (80-. ). **345**:808–811.

**Kumagai, M., Wang, L., and Ueda, S.** (2010). Genetic diversity and evolutionary relationships in genus *Oryza* revealed by using highly variable regions of chloroplast DNA. Gene **462**:44–51.

**Leebens-Mack, J. H., Barker, M. S., Carpenter, E. J., Deyholos, M. K., Gitzendanner, M. A., Graham, S. W., Grosse, I., Li, Z., Melkonian, M., Mirarab, S., et al.** (2019). One thousand plant transcriptomes and the phylogenomics of green plants. Nature **574**:679–685.

**Li, L., Stoeckert, C. J., and Roos, D. S.** (2003). OrthoMCL: Identification of ortholog groups for eukaryotic genomes. Genome Res. **13**:2178–2189.

**Liu, Q., Triplett, J. K., Wen, J., and Peterson, P. M.** (2011). Allotetraploid origin and divergence in Eleusine (Chloridoideae, Poaceae): Evidence from low-copy nuclear gene phylogenies and a plastid gene chronogram. Ann. Bot. **108**:1287–1298.

**Liu, Q., Liu, H., Wen, J., and Peterson, P. M.** (2014). Infrageneric phylogeny and temporal divergence of *Sorghum* (Andropogoneae, Poaceae) based on low-copy nuclear and plastid sequences. PLoS One **9**:e104933.

**Lundgren, M. R., Christin, P. A., Escobar, E. G., Ripley, B. S., Besnard, G., Long, C. M., Hattersley, P. W., Ellis, R. P., Leegood, R. C., and Osborne, C. P.** (2016). Evolutionary implications of C3–C4 intermediates in the grass *Alloteropsis semialata*. Plant Cell Environ. **39**:1874–1885.

**Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Y., et al.** (2012). SOAPdenovo2: An empirically improved memory-efficient short-read de novo assembler. Gigascience **1**:2047-217X-1–18.

**Magallón, S., Gómez-Acevedo, S., Sánchez-Reyes, L. L., and Hernández-Hernández, T.** (2015). A metacalibrated time-tree documents the early rise of flowering plant phylogenetic diversity. New Phytol. **207**:437–453.

**Mai, U., and Mirarab, S.** (2018). TreeShrink: Fast and accurate detection of outlier long branches in collections of phylogenetic trees. BMC Genomics **19**.

**Manchester, S. R.** (2001). Update on the Megafossil Flora of Florissant, Colorado. Proc. Denver Museum Nat. Sci. **4**:137–161.

**Mandel, J. R., Dikow, R. B., Siniscalchi, C. M., Thapa, R., Watson, L. E., and Funk, V. A.** (2019). A fully resolved backbone phylogeny reveals numerous dispersals and explosive diversifications throughout the history of Asteraceae. Proc. Natl. Acad. Sci. U. S. A. **116**:14083–14088.

**Marchant, A. D., and Briggs, B. G.** (2007). Ecdeiocoleaceae and Joinvilleaceae, sisters of Poaceae (Poales): Evidence from *rbcL* and *matK* data. Telopea **11**:437–450.

**Marcussen, T., Sandve, S. R., Heier, L., Spannagl, M., Pfeifer, M., Jakobsen, K. S., Wulff, B., Steuernagel, B., Mayer, K. F. X., Olsen, O. A., et al.** (2014). Ancient hybridizations among the ancestral genomes of bread wheat. Science (80-. ). **345**.

**Mason-Gamer, R. J., Burns, M. M., and Naum, M.** (2010). Reticulate evolutionary history of a complex group of grasses: Phylogeny of *Elymus* StStHH allotetraploids based on three nuclear genes. PLoS One **5**.

**Massidon, W. P., and Maddison, D. R.** (2019). Mesquite: A modular system for evolutionary analysis. Version 3.6. http://www.Mesquiteproject.Org Advance Access published 2019.

**Mathews, S., Spangler, R. E., Mason-Gamer, R. J., and Kellogg, E. A.** (2002). Phylogeny of Andropogoneae inferred from phytochrome B, GBSSI, and NDHF. Int. J. Plant Sci. **163**:441–450.

**Matsumura, H., Xie, Y., Shirakata, S., Inoue, T., Yoshinaga, T., Ueno, Y., Izui, K., and Kai, Y.** (2002). Crystal structures of C4 form maize and quaternary complex of *E. coli* phosphoenolpyruvate carboxylases. Structure **10**:1721–1730.

**McKown, A. D., Moncalvo, J. M., and Dengler, N. G.** (2005). Phylogeny of *Flaveria* (Asteraceae) and inference of C4 photosynthesis evolution. Am. J. Bot. **92**:1911–1928.

**Moreno-Villena, J. J., Dunning, L. T., Osborne, C. P., and Christin, P. A.** (2018). Highly expressed genes are preferentially co-opted for C4 photosynthesis. Mol. Biol. Evol. **35**:94–106.

**Morrone, O., Aagesen, L., Scataglini, M. A., Salariato, D. L., Denham, S. S., Chemisquy, M. A., Sede, S. M., Giussani, L. M., Kellogg, E. A., and Zuloaga, F. O.** (2012). Phylogeny of the Paniceae (Poaceae: Panicoideae): Integrating plastid DNA sequences and morphology into a new classification. Cladistics **28**:333–356.

**Muhaidat, R., Sage, R. F., and Dengler, N. G.** (2007). Diversity of Kranz anatomy and biochemistry in C4 eudicots. Am. J. Bot. **94**:362–381.

**Muller, J.** (1981). Fossil pollen records of extant angiosperms. Bot. Rev. **47**:1–142.

**Nambudiri, E. M. V., Tidwell, W. D., Smith, B. N., and Hebbert, N. P.** (1978). A C4 plant from the Pliocene. Nature **276**:816–817.

**Newman, Z. R., Young, J. M., Ingolia, N. T., and Barton, G. M.** (2016). Differences in codon bias and GC content contribute to the balanced expression of TLR7 and TLR9. Proc. Natl. Acad. Sci. U. S. A. **113**:E1362–E1371.

**Nguyen, L. T., Schmidt, H. A., Von Haeseler, A., and Minh, B. Q.** (2015). IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. **32**:268–274.

**Nobel, P. S., and Hartsock, T. L.** (1986). Leaf and stem CO2 uptake in the three subfamilies of the Cactaceae. Plant Physiol. **80**:913–917.

**O'Leary, B., Fedosejevs, E. T., Hill, A. T., Bettridge, J., Park, J., Rao, S. K., Leach, C. A., and Plaxton, W. C.** (2011). Tissue-specific expression and post-translational modifications of plant-and bacterial-type phosphoenolpyruvate carboxylase isozymes of the castor oil plant, *Ricinus communis* L. J. Exp. Bot. **62**:5485–5495.

**Paulus, J. K., Schlieper, D., and Groth, G.** (2013). Greater efficiency of photosynthetic carbon fixation due to single amino-acid substitution. Nat. Commun. **4: 1518**.

**Perreta, M. G., Ramos, J. C., and Vegetti, A. C.** (2009). Development and structure of the grass inflorescence. Bot. Rev. **75**:377–396.

**Peterson, P. M., Romaschenko, K., and Johnson, G.** (2010). A classification of the Chloridoideae (Poaceae) based on multi-gene phylogenetic trees. Mol. Phylogenet. Evol. **55**:580–598.

**Peterson, P. M., Romaschenko, K., Barker, N. P., and Linder, H. P.** (2011). Centropodieae and *Ellisochloa*, a new tribe and genus in Chloridoideae (Poaceae). Taxon **60**:1113–1122.

**Peterson, P. M., Romaschenko, K., and Arrieta, Y. H.** (2014). A molecular phylogeny and classification of the Cteniinae, Farragininae, Gouiniinae, Gymnopogoninae, Perotidinae, and Trichoneurinae (Poaceae: Chloridoideae: Cynodonteae). Taxon **63**:275–286.

**Peterson, P. M., Romaschenko, K., and Arrieta, Y. H.** (2015). Phylogeny and subgeneric classification of *Bouteloua* with a new species, *B. herrera-arrietae* (Poaceae: Chloridoideae: Cynodonteae: Boutelouinae). J. Syst. Evol. **53**:351–366.

**Philippe, H., Brinkmann, H., Lavrov, D. V., Littlewood, D. T. J., Manuel, M., Wörheide, G., and Baurain, D.** (2011). Resolving difficult phylogenetic questions: Why more sequences are not enough. PLoS Biol. **9:e1000602**.

**Prasad, V., Strömberg, C. A. E., Leaché, A. D., Samant, B., Patnaik, R., Tang, L., Mohabey, D. M., Ge, S., and Sahni, A.** (2011). Late Cretaceous origin of the rice tribe provides evidence for early diversification in Poaceae. Nat. Commun. **2:480**.

**Rivoal, J., Trzos, S., Gage, D. A., Plaxton, W. C., and Turpin, D. H.** (2001). Two unrelated phosphoenolpyruvate carboxylase polypeptides physically interact in the high molecular mass isoforms of this enzyme in the unicellular green alga *Selenastrum minutum*. J. Biol. Chem. **276**:12588–12597.

**Saarela, J. M., Burke, S. V., Wysocki, W. P., Barrett, M. D., Clark, L. G., Craine, J. M., Peterson, P. M., Soreng, R. J., Vorontsova, M. S., and Duvall, M. R.** (2018). A 250 plastome phylogeny of the grass family (Poaceae): Topological support under different data partitions. PeerJ **2018**.

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

**Sage, R. F.** (2016). A portrait of the C4 photosynthetic family on the 50th anniversary of its discovery: Species number, evolutionary lineages, and Hall of Fame. J. Exp. Bot. **67**:4039–4056.

**Sage, R. F., Monson, R. K., Ehleringer, J. R., Adachi, S., and Pearcy, R. W.** (2018). Some like it hot: the physiological ecology of C4 plant evolution. Oecologia **187**:941–966.

**Sánchez-Ken, J. G., and Clark, L. G.** (2001). Gynerieae, a new neotropical tribe of grasses (Poaceae). Novon **11**:350–352.

**Sánchez-Ken, J. G., Clark, L. G., Kellogg, E. A., and Kay, E. E.** (2007). Reinstatement and Emendation of Subfamily Micrairoideae (Poaceae). Syst. Bot. **32**:71–80.

**Sánchez, R., and Cejudo, F. J.** (2003). Identification and expression analysis of a gene encoding a bacterial-type phosphoenolpyruvate carboxylase from *Arabidopsis* and rice. Plant Physiol. Advance Access published 2003, doi:10.1104/pp.102.019653.

**Sayyari, E., and Mirarab, S.** (2016). Fast Coalescent-Based Computation of Local Branch Support from Quartet Frequencies. Mol. Biol. Evol. **33**:1654–1668.

**Schlüter, U., and Weber, A. P. M.** (2020). Regulation and evolution of C4 photosynthesis. Annu. Rev. Plant Biol. 2020 **71**:183–215.

**Schneider, J., Winterfeld, G., Hoffmann, M. H., and RöSer, M.** (2011). Duthieeae, a new tribe of grasses (Poaceae) identified among the early diverging lineages of subfamily Pooideae: molecular phylogenetics, morphological delineation, cytogenetics and biogeography. Syst. Biodivers. **9**:27–44.

**Schubert, M., Marcussen, T., Meseguer, A. S., and Fjellheim, S.** (2019). The grass subfamily Pooideae: Cretaceous–Palaeocene origin and climate-driven Cenozoic diversification. Glob. Ecol. Biogeogr. **28**:1168–1182.

**Sedelnikova, O. V., Hughes, T. E., and Langdale, J. A.** (2018). Understanding the genetic basis of C4 Kranz anatomy with a view to engineering C3 crops. Annu. Rev. Genet. **52**:249–270.

**Serres-Giardi, L., Belkhir, K., David, J., and Glémin, S.** (2012). Patterns and evolution of nucleotide landscapes in seed plants. Plant Cell **24**:1379–1397.

**Sharp, P. M., and Li, W. H.** (1987). The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. **15**:1281–1295.

**Shi, G., Grimaldi, D. A., Harlow, G. E., Wang, J., Wang, J., Yang, M., Lei, W., Li, Q., and Li, X.** (2012). Age constraint on Burmese amber based on U-Pb dating of zircons. Cretac. Res. **37**:155–163.

**Shimodaira, H., and Hasegawa, M.** (1999). Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol. Biol. Evol. **16**:1114–1116.

**Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., et al.** (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. **7**.

**Simon, B. K.** (2007). GrassWorld - Interactive key and information system of world grasses. Kew Bull. **62**:475–484.

**Sinha, N. R., and Kellogg, E. A.** (1996). Parallelism and diversity in multiple origins of C4 photosynthesis in the grass family. Am. J. Bot. **83**:1458–1470.

**Skendzic, E., Columbus, T., and Cerros-Tlatilpa, R.** (2007). Phylogenetics of Andropogoneae (Poaceae: Panicoideae) Based on Nuclear Ribosomal Internal Transcribed Spacer and Chloroplast *trnL–F* Sequences. Aliso **23**:530–544.

**Smith, B. N., and Epstein, S.** (1971). Two categories of $^{13}C^{12}/C$ ratios for higher plants. Plant Physiol. **47**:380–384.

**Smith, S. A., and O'Meara, B. C.** (2012). TreePL: Divergence time estimation using penalized likelihood for large phylogenies. Bioinformatics **28**:2689–2690.

**Soderstrom, T. R.** (1981). Some Evolutionary Trends in the Bambusoideae (Poaceae). Ann. Missouri Bot. Gard. **68**:15–47.

**Soreng, R. J., Peterson, P. M., Romaschenko, K., Davidse, G., Zuloaga, F. O., Judziewicz, E. J., Filgueiras, T. S., Davis, J. I., and Morrone, O.** (2015). A worldwide phylogenetic classification of the Poaceae (Gramineae). J. Syst. Evol. **53**:117–137.

**Soreng, R. J., Peterson, P. M., Romaschenko, K., Davidse, G., Teisher, J. K., Clark, L. G., Barberá, P., Gillespie, L. J., and Zuloaga, F. O.** (2017). A worldwide phylogenetic classification of the Poaceae (Gramineae) II: an update and a comparison of two 2015 classifications. J. Syst. Evol. **55**:259–290.

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

**Stöver, B. C., and Müller, K. F.** (2010). TreeGraph 2: Combining and visualizing evidence from different phylogenetic analyses. BMC Bioinformatics Advance Access published 2010, doi:10.1186/1471-2105-11-7.

**Strömberg, C. A. E.** (2005). Decoupled taxonomic radiation and ecological expansion of open-habitat grasses in the Cenozoic of North America. Proc. Natl. Acad. Sci. U. S. A. **102**:11980–11984.

**Struck, T. H.** (2014). Trespex-detection of misleading signal in phylogenetic reconstructions based on tree information. Evol. Bioinforma. **10**:51–67.

**Tang, L., Zou, X. hui, Achoundong, G., Potgieter, C., Second, G., Zhang, D. yong, and Ge, S.** (2010a). Phylogeny and biogeography of the rice tribe (Oryzeae): Evidence from combined analysis of 20 chloroplast fragments. Mol. Phylogenet. Evol. **54**:266–277.

**Tang, H., Bowers, J. E., Wang, X., and Paterson, A. H.** (2010b). Angiosperm genome comparisons reveal early polyploidy in the monocot lineage. Proc. Natl. Acad. Sci. U. S. A. **107**:472–477.

**Teerawatananon, A., Jacobs, S. W. L., and Hodkinson, T. R.** (2011). Phylogenetics of Panicoideae (Poaceae) based on chloroplast and nuclear DNA sequences. Telopea **13**:115–142.

**Teisher, J. K., McKain, M. R., Schaal, B. A., and Kellogg, E. A.** (2017). Polyphyly of Arundinoideae (Poaceae) and evolution of the twisted geniculate lemma awn. Ann. Bot. **120**:725–738.

**Terada, K., and Izui, K.** (1991). Site-directed mutagenesis of the conserved histidine residue of phosphoenolpyruvate carboxylase: His 138 is essential for the second partial reaction. Eur. J. Biochem. **202**:797–803.

**Timilsena PR, Wafula EK, Barrett CF, Ayyampalayam S, McNeal JR, Rentsch JD, McKain MR, Heyduk K, Harkess A, Villegente M, Conran JG, Illing N, Fogliani B, Ane´ C, Pires JC, Davis JI, Zomlefer WB, Stevenson DW, Graham SW, Givnish TJ, L.-M. J. and dePamphilis C.** (2022). Phylogenomic resolution of order- and family-level monocot relationships using 602 single-copy nuclear genes and 1375 BUSCO genes. Front. Plant Sci. **13:876779**.

**Tregunna, E. B., Smith, B. N., Berry, J. A., and Downton, W. J. S.** (1970). Some methods for studying the photosynthetic taxonomy of the angiosperms. Can. J. Bot. **48**:1209–1214.

**Triplett, J. K., Clark, L. G., Fisher, A. E., and Wen, J.** (2014). Independent allopolyploidization events preceded speciation in the temperate and tropical woody bamboos. New Phytol. **204**:66–73.

**Tzvelev, N. N.** (1989). The system of grasses (Poaceae) and their evolution. Bot. Rev. **55**:141–204.

**Van Leuven, J. T., and McCutcheon, J. P.** (2012). An AT mutational bias in the tiny GC-rich endosymbiont genome of *Hodgkinia*. Genome Biol. Evol. **4**:24–27.

**Vegetti, A., and Anton, A. M.** (1995). Some evolution trends in the inflorescence of Poaceae. Flora **190**:225–228.

**Vicentini, A., Barber, J. C., Aliscioni, S. S., Giussani, L. M., and Kellogg, E. A.** (2008). The age of the grasses and clusters of origins of C4 photosynthesis. Glob. Chang. Biol. **14**:2963–2977.

**Vogel, J. P., Garvin, D. F., Mockler, T. C., Schmutz, J., Rokhsar, D., Bevan, M. W., Barry, K., Lucas, S., Harmon-Smith, M., Lail, K., et al.** (2010). Genome sequencing and analysis of the model grass *Brachypodium distachyon*. Nature **463**:763–768.

**Voznesenskaya, E. V., Franceschi, V. R., Kiirats, O., Freitag, H., and Edwards, G. E.** (2001). Kranz anatomy is not essential for terrestrial C4 plant photosynthesis. Nature Advance Access published 2001, doi:10.1038/35107073.

**Walther, H., and Kvaček, Z.** (2007). Early Oligocene flora of Seifhennersdorf (Saxony). Acta Musei Natl. Pragae **63**:85–174.

**Wang, N., Zhong, X., Cong, Y., Wang, T., Yang, S., Li, Y., and Gai, J.** (2016). Genome-wide analysis of phosphoenolpyruvate carboxylase gene family and their response to abiotic stresses in soybean. Sci. Rep. **6**:38448.

**Washburn, J. D., Schnable, J. C., Davidse, G., and Pires, J. C.** (2015). Phylogeny and photosynthesis of the grass tribe Paniceae. Am. J. Bot. **102**:1493–1505.

**Watson, L., and Dallwitz, M. J.** (1992).DELTA – Description Language for Taxonomy. https://www.delta-intkey.com/grass/index.htm.

**Wei, F., Coe, E., Nelson, W., Bharti, A. K., Engler, F., Butler, E., Kim, H. R., Goicoechea, J. L., Chen, M., Lee, S., et al.** (2007). Physical and genetic structure of the maize genome reflects its complex evolutionary history. PLoS Genet. **3**:1254–1263.

**Welker, C. A. D., McKain, M. R., Estep, M. C., Pasquet, R. S., Chipabika, G., Pallangyo, B., and Kellogg, E. A.** (2020). Phylogenomics enables biogeographic analysis and a new subtribal classification of Andropogoneae (Poaceae—Panicoideae). J. Syst. Evol. **58**:1003–1030.

**Whistler, D. P., Tedford, R. H., Takeuchi, G. T., Wang, X., Tseng, Z. J., and Perkins, M. E.** (2009). Revised Miocene biostratigraphy and biochronology of the Dove Spring Formation, Mojave Desert, California. Pap. Geol. Vertebr. Paleontol. Biostratigraphy Honor Michael O. Woodburne Advance Access published 2009.

**Wickett, N. J., Mirarab, S., Nguyen, N., Warnow, T., Carpenter, E., Matasci, N., Ayyampalayam, S., Barker, M. S., Burleigh, J. G., Gitzendanner, M. A., et al.** (2014). Phylotranscriptomic analysis of the origin and early diversification of land plants. Proc. Natl. Acad. Sci. U. S. A. **111**:E4859–E4868.

**Winter, K., Wallace, B. J., Stocker, G. C., and Roksandic, Z.** (1983). Crassulacean acid metabolism in australian vascular epiphytes and some related species. Oecologia **57**:129–141.

**Wu, Y., You, H. L., and Li, X. Q.** (2018). Dinosaur-associated Poaceae epidermis and phytoliths from the Early Cretaceous of China. Natl. Sci. Rev. **5**:721–727.

**Wysocki, W. P., Clark, L. G., Attigala, L., Ruiz-Sanchez, E., and Duvall, M. R.** (2015). Evolution of the bamboos (Bambusoideae; Poaceae): A full plastome phylogenomic analysis. BMC Evol. Biol. **15:50**.

**Wysocki, W. P., Ruiz-Sanchez, E., Yin, Y., and Duvall, M. R.** (2016). The floral transcriptomes of four bamboo species (Bambusoideae; Poaceae): support for common ancestry among woody bamboos. BMC Genomics Advance Access published 2016, doi:10.1186/s12864-016-2707-1.

**Xiang, Y., Huang, C. H., Hu, Y., Wen, J., Li, S., Yi, T., Chen, H., Xiang, J., and Ma, H.** (2017). Evolution of Rosaceae fruit types based on nuclear phylogeny in the context of geological times and genome duplication. Mol. Biol. Evol. **34**:262–281.

**Yakovchuk, P., Protozanova, E., and Frank-Kamenetskii, M. D.** (2006). Base-stacking and base-pairing contributions into thermal stability of the DNA double helix. Nucleic Acids Res. **34**:564–574.

**Yang, Y., Moore, M. J., Brockington, S. F., Mikenas, J., Olivieri, J., Walker, J. F., and Smith, S. A.** (2018). Improved transcriptome sampling pinpoints 26 ancient and more recent polyploidy events in Caryophyllales, including two allopolyploidy events. New Phytol. **217**:855–870.

**Zeng, L., Zhang, Q., Sun, R., Kong, H., Zhang, N., and Ma, H.** (2014). Resolution of deep angiosperm phylogeny using conserved nuclear genes and estimates of early divergence times. Nat. Commun. **5**:4956.

**Zhang, J., Lu, H., Gu, W., Wu, N., Zhou, K., Hu, Y., Xin, Y., and Wang, C.** (2012a). Early mixed farming of millet and rice 7800 years ago in the middle yellow river region, China. PLoS One **7**.

**Zhang, N., Zeng, L., Shan, H., and Ma, H.** (2012b). Highly conserved low-copy nuclear genes as effective markers for phylogenetic analyses in angiosperms. New Phytol. **195**:923–937.

**Zhang, Y. X., Ma, P. F., and Li, D. Z.** (2018). A new genus of temperate woody bamboos (Poaceae, Bambusoideae, Arundinarieae) from a limestone montane area of China. PhytoKeys **109**:67–76.

**Zhang, J., Zhang, Q., Li, L., Tang, H., Zhang, Q., Chen, Y., Arrow, J., Zhang, X., Wang, A., Miao, C., et al.** (2019). Recent polyploidization events in three Saccharum founding species. Plant Biotechnol. J. **17**:264–274.

**Zhang, Y. X., Guo, C., and Li, D. Z.** (2020). A new subtribal classification of Arundinarieae (Poaceae, Bambusoideae) with the description of a new genus. Plant Divers. **42**:127–134.

**Zhang, L., Zhu, X., Zhao, Y., Guo, J., Zhang, T., Huang, W., Huang, J., Hu, Y., Huang, C.-H., and Ma, H.** (2022). Phylotranscriptomics resolves the phylogeny of Pooideae and uncovers factors for their adaptive evolution. Mol. Biol. Evol. **39**.

**Zhao, L., Zhang, N., Ma, P. F., Liu, Q., Li, D. Z., and Guo, Z. H.** (2013). Phylogenomic analyses of nuclear genes reveal the evolutionary relationships within the BEP clade and the evidence of positive selection in Poaceae. PLoS One **8**:e64642.

**Zhao, Y., Zhang, R., Jiang, K. W., Qi, J., Hu, Y., Guo, J., Zhu, R., Zhang, T., Egan, A. N., Yi, T. S., et al.** (2021). Nuclear phylotranscriptomics and phylogenomics support numerous polyploidization events and hypotheses for the evolution of rhizobial nitrogen-fixing symbiosis in Fabaceae. Mol. Plant **14**:748–773.

**Zhu, Q., and Ge, S.** (2005). Phylogenetic relationships among A-genome species of the genus *Oryza* revealed by intron sequences of four nuclear genes. New Phytol. **167**:249–265.

**Zhu, J., He, F., Hu, S., and Yu, J.** (2008). On the nature of human housekeeping genes. Trends Genet. **24**:481–484.

**Zou, X.-H., Zhang, F.-M., Zhang, J.-G., Zang, L.-L., Tang, L., Wang, J., Sang, T., and Ge, S.** (2008). Analysis of 142 genes resolves the rapid diversification of the rice genus. Genome Biol. **9**:R49.

**Zou, X. H., Yang, Z., Doyle, J. J., and Ge, S.** (2013). Multilocus estimation of divergence times and ancestral effective population sizes of *Oryza* species and implications for the rapid diversification of the genus. New Phytol. **198**:1155–1164.

**Zucol, A. F., Brea, M., and Bellosi, E. S.** (2010). Phytolith studies in Gran Barranca (central Patagonia, Argentina): the middle-late Eocene. In *The Paleontology of Gran Barranca:*

*Evolution and Environmental Change through the Middle Cenozoic of Patagonia*, pp. 317–340.

# VITA

## Weichen Huang

### Education

The Pennsylvania State University, 8/2017-present, PhD candidate in biology, expected graduation: May 2023

Zhejiang University, 09/2013-06/2017, major: biotechnology, degree: Bachelor of Science

### Research Projects

08/2017-present, Poaceae phylogeny based on low-copy nuclear genes and the evolution of $C_4$ photosynthesis.

08/2016-05/2017, The evolutionary adaptation of body size of animals on land-bridge islands in Thousand Island Lake (National Natural Science Foundation Projects of China).

05/2015-06/2015, North Carolina State University-Zhejiang University joint course: plant resources, ecology & culture of eastern China.

### Publications

**Huang, W., Zhang, L., Columbus, J. T., Hu, Y., Zhao, Y., Tang, L., Guo, Z., Chen, W., McKain, M., Bartlett and Ma, H.** (2022). A well-supported nuclear phylogeny of Poaceae and implications for the evolution of $C_4$ photosynthesis. Mol. Plant **15**: 755–777.

**Zhang, L., Zhu, X., Zhao, Y., Guo, J., Zhang, T., Huang, W., Huang, J., Hu, Y., Huang, C. H., and Ma, H.** (2022). Phylotranscriptomics resolves the phylogeny of Pooideae and uncovers factors for their adaptive evolution. Mol. Biol. Evol. **39:** msac026.