2016

A bioinformatic study on the evolution of bamboos and other graminoid poales

William P. Wysocki

ABSTRACT

A BIOINFORMATIC STUDY ON THE EVOLUTION OF BAMBOOS AND OTHER GRAMINOID POALES

William P. Wysocki, Ph.D.
Department of Biological Sciences
Northern Illinois University, 2016
Melvin R. Duvall, Director

For the past decade, next-generation sequencing (NGS) has been used to undertake genomics-scale projects in molecular biology. This method of sequencing involves randomly fragmenting a sample of nucleic acid and randomly generating millions of short reads. The large number of reads moves the assembly and other analyses to be performed by computer algorithms. As well as sequencing full chromosomes, full RNA extracts can be sequenced to determine levels of gene expression and exon boundaries.

In this study, NGS is used to examine the evolution of bamboos (Bambusoideae), which are a subfamily of grasses (Poaceae). Bamboos are divided into two main phenotypes: woody and herbaceous. Woody bamboos are characterized by lignified culms, bisexual florets and undergo gregarious monocarpy, while herbaceous bamboos have less lignified shoots, unisexual florets and flower annually.

The evolution of this subfamily was examined in a phylogenomic framework using full chloroplast genome (plastome) sequences. First, a set of methods for plastome assembly was developed using automatable scripts and accuracy-testing steps. These methods were then used to generate full plastomes from bamboos. Full plastomes were then analyzed phylogenomically under a maximum-likelihood and Bayesian framework. This analysis

revealed paraphyly between the temperate woody and tropical woody clades. Previously established tribal and subtribal relationships were also confirmed.

Nuclear transcripts were assembled from four bamboo species by sequencing RNA from floral tissue using NGS. Two assembly methods were performed: a de novo assembly used only overlapping reads to produce transcripts and a reference-based assembly used a previously sequenced nuclear genome from a bamboo to place reads and assemble them into transcripts. After quality assessment, the reference-based produced about double the number of transcripts than then de novo assembly, but may be biased toward the number of nuclear genes found in the reference genome. Floral development transcripts were examined as gene trees. Several preliminary correlations to phenotype were determined using putative gene function. The nuclear transcripts were then used to perform a phylogenetic analysis. After 3,878 transcripts were determined to be single-copy, they were concatenated and analyzed under a maximum-likelihood framework. Each of the 3,878 transcripts were also analyzed individually. This analysis strongly supported a monophyletic relationship between the woody bamboos, which contrasted with the plastome analysis.

Finally, a plastome was generated for a near-grass *Joinvillea ascendens* (Joinvilleaceae). The plastome was generated using completely de novo methods and was found to have undergone two large-scale inversions. These inversions occurred after two other large-scale inversions in the Poaceae-Joinvilleaceae lineage, which had been previously documented using PCR amplification. The two previously documented inversions were verified by comparing grasses to a plastome with an ancestral gene arrangement. Several gene and intron losses were also verified in grasses.

NORTHERN ILLINOIS UNIVERSITY
DE KALB, ILLINOIS

MAY 2016

A BIOINFORMATIC STUDY ON THE EVOLUTION OF BAMBOOS AND OTHER GRAMINOID POALES

BY

WILLIAM P WYSOCKI

A DISSERTATION SUBMITTED TO THE GRADUATE SCHOOL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE

DOCTOR OF PHILOSOPHY

DEPARTMENT OF BIOLOGICAL SCIENCES

Doctoral Director:
Melvin R. Duvall

ACKNOWLEDGEMENTS

I thank Dr. Mel Duvall for his encouragement and support these past four years. I also thank Dr. Yanbin Yin, Dr. Lynn Clark, Dr. Wes Swingley and Dr. Tom Sims for their participation in my dissertation committee. I would also like to thank Dr. Eduardo Ruiz-Sanchez for his help in getting around Mexico and locating flowering bamboo populations. I thank Dr. Paul Peterson for allowing me to access the collection at the U.S. National Herbarium.

I also would like to thank all of my collaborators, especially Scot Kelchner, Joe Cotton, Sean Burke, Lauren Orton, Jane Digiovanni, Jeff Saarela, Nick Barber, Jerry Davis, Lakshmi Attigala, J. Chris Pires and Patrick Edger. I am thankful for the work performed by the undergraduates who helped in data collection and assembly including Keith Murrell, LeRoy Reinke, Taylor Nicholas, Jakob Stricker, Ben Walter, Lewis Schrank and Olivia East. I would also like to thank Collin Jaeger for providing a dissertation template.

TABLE OF CONTENTS

Chapter Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

Figure Page

PREFACE

Grasses (Poaceae) are a family of monocots and the fifth largest of all plant families. In addition to being an essential food source for humans worldwide, grasses are important ecologically as a food source and habitat for other organisms. Bambusoideae is one of the twelve grass subfamilies, which falls into the Bambusoideae-Oryzoideae-Pooideae (BOP) clade and is phylogenetically sister to Pooideae (Wu and Ge, 2009). Species within Bambusoideae are colloquially known as bamboos and phenotypically fall into two categories: woody and herbaceous bamboos. Woody bamboos form hard lignified culms, which can be used as building or crafting material. These bamboos can generate forest-like biomes and serve as a shady habitat. In addition to their shoot characteristics, woody bamboos are also united by their floral morphologies and phenological patterns. Woody bamboos possess bisexual flowers and undergo synchronized flowering events, which occur in cycles that last from several to 120 years, followed by a die-off (http://www.eeob.iastate.edu/research/bamboo/bamboo.html). The herbaceous bamboos are inconspicuous with lower levels of lignification in their shoots. Herbaceous bamboos are typically polycarpic perennials, which have unisexual flowers and undergo annual flowering.

Phylogenetic studies, as inferred from multi-locus analyses (Bouchenak-Khelladi et al., 2008; Kelchner et al., 2013), have identified deep divergence patterns in the bamboo tree of life. However, additional information is required to increase resolution and support for many of the relationships; especially among the woody bamboo clades. Inferring phylogeny through the use of large-scale genomics data, phylogenomics, increases the number of

informative characters and uncovers genomic features that would be left obscured by the limited sequencing of certain regions. Phylogenomics using full chloroplast genomes (plastomes), which can utilize all of the uniparentally-inherited phylogenetic signal present in the plastid compartment, is particularly useful in grasses. Intrafamilial studies in grasses using full plastomes can be accomplished as the entire chromosome is alignable due to conserved gene order. Phylogenomics can also be applied to biparentally-inherited nuclear genes by using full sets of expressed transcripts (transcriptomes) for phylogenetic analysis. Phylogenomics can also be applied to non-alignable full chromosomes by examining the large scale changes that occurred between them. Examining both types of markers is necessary to clarify evolutionary patterns, especially in grasses, which hybridize readily.

Next-generation sequencing:

In the past fifteen years, DNA sequencing evolved significantly from the traditional Sanger methods, which used targeted one PCR product in a dideoxynucleotide chain termination reaction to produce one read of 1-2 kbp. This evolution, called next-generation sequencing (NGS), uses a shotgun sequencing approach to randomly produce short reads (50--500 bp) that can range in number from tens of thousands to hundreds of millions. The large number of reads necessitates a shift of the assembly aspect to computers. Short reads can be assembled into contiguous sequences (contigs) using either sequence similarity to a reference genome to place reads, or de novo methods, which rely on identifying the overlapping regions shared between reads. Location of sequencing products within contigs is also shifted to bioinformatics methods due to the random-targeting of NGS and large amount of data produced.

The use of NGS has increased the potential for phylogenomic studies to be performed at increasingly larger scales. Prior to the advent of NGS, only a few dozen grass plastomes had been sequenced, but the implementation of NGS has increased the number of sequenced plastomes to over 300 (~150 in GenBank, over 150 unpublished by the Duvall laboratory). This increase has occurred largely because plastome assembly has shifted to computer-based methods, which can be automated through the use of high-throughput scripts and pipelines. The potential to undergo projects in nuclear phylogenomics has also increased with NGS. While sequencing a full nuclear genome requires deep sequencing and extensive processing power, a representation of thousands of unique nuclear transcripts can be assembled using RNA-Seq. After an assessment of orthology is performed, nuclear transcripts can then be used to infer phylogeny.

Dissertation outline:

This dissertation uses bioinformatics to assemble NGS products from bamboos and other graminid Poales and analyze them phylogenomically (in this case at the whole plastome-scale) and transcriptomically. Most sequencing was performed on an Illumina platform, which produced millions to tens of millions of ~100 bp reads per sample. The plastome assembly portion of the dissertation used single-end reads, which produce one read per sequenced fragment. The RNA-Seq portion used paired-end reads, which produce one read from one end of each fragment and one additional read from the other end. This provides additional information on the relative position of each read.

The first chapter seeks to justify the use of the relatively young NGS techniques and to produce an accurate and efficient bioinformatics protocol for the sequencing, assembly

and annotation of grass plastomes. Justification of NGS was accomplished by sequencing the plastome of one species (*Neyraudia reynaudiana*) in duplicate using traditional Sanger technologies as well as NGS and comparing the results. Three additional plastomes were assembled using NGS and compared to previously assembled plastomes from congeneric species, which were expected to be nearly identical if the assembly methods were accurate. Using these comparisons, a protocol was developed by selecting software packages for de novo assembly and producing custom scripts that scaffolded assembled contigs. A method for accurate assembly verification using raw reads was also developed. The methods and protocols developed in the chapter were used for all plastome assembly performed in the rest of this dissertation.

Chapters 2 and 3 focus on bamboo phylogeny using full plastomes. A robust phylogeny was produced in the second chapter using 31 plastomes that represent the three main bamboo tribes. Several novel bamboo-specific plastid features were elucidated including the first documented mitochondrion-to-plastid transfer in the monocots. The third chapter expanded on this study and added 53 newly sequenced plastomes to the phylogeny. In addition to a near complete subtribal representation, larger genera were then thoroughly represented. Several previously-sequenced plastomes were assembled independently, in duplicate, to verify the phylogenetic position of each.

The fourth chapter focuses mainly on bamboo nuclear transcriptomics. Floral transcriptomes from two representatives of Bambuseae, one representative from Arundinarieae, and one representative from Olyreae were generated using RNA-Seq. A de novo assembly and reference-based assembly, using the draft genome from *Phyllostachys*

heterocycla, were produced and compared according to content. Because of the differences between the woody and herbaceous bamboos in floral structure and timing, the evolutionary history of genes that control these characteristics was explored. Additionally, transcripts were extracted according to orthology and used to perform a large-scale nuclear phylogenomics analysis using over 3,700 loci.

The fifth chapter explores the general structure of the grass plastome by comparing it to the newly-sequenced plastome from *Joinvillea ascendens*, a grass-like member of the Poales. This study elucidates the large-scale inversion events that occurred in the lineages leading to the evolution of *Joinvillea* and the grasses. Evidence of events that were previously detected using traditional PCR-based methods were also located and verified. These lineages are compared to the *Typha latifolia* plastome, due to its shared gene order with the 'general' eudicot plastome structure.

CHAPTER 1

A MULTI-STEP COMPARISON OF SHORT-READ FULL PLASTOME SEQUENCE ASSEMBLY METHODS IN GRASSES

ABSTRACT

Technological advances have allowed phylogenomic studies of plants, such as full chloroplast genome (plastome) analysis, to become increasingly popular and economically feasible. Although next-generation short-read sequencing allows for full plastomes to be sequenced relatively rapidly, it requires additional attention using software to assemble these reads into comprehensive sequences. Here we compared the use of three de novo assemblers combined with three contig assembly methods. Seven plastome sequences were analyzed. Three of these were Sanger-sequenced. The other four were assembled from short, single-end read files generated from next-generation libraries. These plastomes represented a total of six grass species (Poaceae), one of which was sequenced in duplicate by the two methods to allow direct comparisons for accuracy. Enumeration of missing sequence and ambiguities allowed for assessment of completeness and efficiency. All methods that used de Bruijn-based de novo assemblers were shown to produce assemblies comparable to the Sanger-sequenced plastomes but were not equally efficient in producing complete plastomes. Contig assembly methods that utilized automatable and repeatable processes were generally more efficient and advantageous when applied to larger scale projects with many full plastomes. However, contig assembly methods that were less automatable and required more manual

attention did show utility in determining plastomes with lower read depth that could not be assembled with automatable procedures. Although the methods here were used exclusively to generate grass plastomes, these could be applied to other taxonomic groups if previously sequenced plastomes were available.

INTRODUCTION

Systematic studies of land plants have advanced in the context of molecular phylogenetic research. Access to next-generation sequencing (NGS) methods has given plant systematists the ability to analyze genome-scale data for large numbers of terminal taxa to investigate evolutionary relationships. Taxa within the grass family (Poaceae) are of particular interest due to the economic importance of cereal grains, their use in functional genomic studies (Botiri et al., 2008) and their complex evolutionary history. These phylogenomic studies—which are genome-scale phylogenetic studies—have been shown to offer improved resolution, stronger support, and allow for more confident estimates of divergence dates. In this context, complete chloroplast genomes (plastomes) have been leveraged to show fine-scale relationships, document microstructural events, and offer explanations for historical biogeographic patterns (Cronn et al., 2008; Parks et al., 2009, 2012; Burke et al., 2012, 2014). Intergeneric studies show broader evolutionary patterns even in taxa with slowly evolving plastomes, indicate patterns of genome evolution, and show lineage-specific rate variations (Zhang, Y.J. et al., 2011; Hand et al., 2013). Studies within families have resolved formerly intractable phylogenetic issues and revealed readily interpretable patterns of molecular evolution (Leseberg and Duvall, 2009; Duvall et al., 2010; Wu & Ge, 2012).

Growing use of complete plastomes by plant systematists does not reflect uniform methodological choices for determining and assembling these data. In part, this is due to the rapid changes of the NGS technologies, with periodic advances often accompanied by increases in read length, which directly impact the efficiency and accuracy of assembly. The development of new software tools for assembly is another factor. However, these do not entirely account for the diversity of methods of plastome assembly employed in published reports. Although there are numerous comparative studies of different methods of assembling large genomes (Lin et al., 2011; Zhang, W. et al., 2011; Liu et al., 2012; and many others), less is documented on assembling complete plastomes (Steele et al., 2012). Plastome assembly presents challenges such as a large inverted repeat region, mitogenomes with similar sequences that are likely intermingled in the same pool of reads, and AT-richness, which introduces periodic regions of low sequence complexity. Plastomes also present unique opportunities for phylogenomic analysis, foremost of which are their highly conserved sequences and structures. Major structural changes have been occasionally well documented at intergeneric or interfamilial levels (e.g., Doyle et al., 1992; Cosner et al., 2004). However, plastomes from monocot congeners such as *Acorus americanus* (Raf.) Raf. (NC010093) and *A. calamus* L. (NC007407) and those from *Oryza sativa* L. "Japonica Group" (NC001320) and "Indica Group" (NC008155), *O. meridionalis* N.Q.Ng (NC016927) and *O. nivara* S.D.Shartma & Shastry (NC005973) show 99.54% and 99.49% nucleotide identities in alignments of these plastomes respectively (Wysocki, unpub. comparisons).

In existing studies of complete plastomes, different combinations of assembly methods have been used. Prior to assembly, NGS reads may be trimmed based on sequence

quality (although see Paszkiewicz & Studholme, 2010 for a contrasting view). Reads may also be filtered by comparison against published plastomes, discarding those that fail to meet a threshold nucleotide identity (e.g., Hand et al., 2013). The risk here is that low frequency events caused by intergenomic recombinations within a species may be missed. Assemblies may be accompanied by a reference-guided step, where de novo assembled contigs are aligned to the plastome of a closely related species (Zhang, Y.J. et al., 2011), or even where an intermediate pseudoreference is created that is chimeric between the de novo sequences and the reference (Whittall et al., 2010). Some Sanger-sequencing of plastome fragments may also be performed to close gaps in the assembly (Hand et al., 2013), verify assembled sequences in mutation hotspots (Whittall et al., 2010) or to identify the boundaries of the major inverted repeats (Zhang, Y.J. et al., 2011).

A robust test for accuracy of plastome assembly is the comparison of duplicated sequences from the same plant using two different methods, such as Sanger and Illumina. We sequenced a plastome from *Neyraudia reynaudiana* (Kunth) Keng ex Hitchc. in duplicate using Sanger and NGS technology. Additionally we used Sanger-sequenced plastomes of two other species (*Arundinaria gigantea* (Walt.) Muhl., *Pharus latifolius* L.) along with NGS sequenced plastomes from closely related congeners of these species to perform a somewhat less stringent test, similar to the strategy employed by Steele et al. (2012). This is justified because of the high identities of plastomes between grass and other monocot congeners (see above).

The study species represent three major lineages of grasses. From the PACMAD clade (acronym abbreviates the subfamilial membership for Panicoideae, Arundinoideae,

Chloridoideae, Micrairoideae, Aristidoideae, and Danthonioideae) the chloridoid *Neyraudia reynaudiana* was selected and compared against itself for the two sequencing methods. From the BEP clade (acronym for: Bambusoideae, Ehrhartoideae, and Pooideae) the New World bambusoid species (following Clark and Triplett, 1993) *Arundinaria gigantea* and both of its congeners, *A. tecta* (Walt.) Muhl. and *A. appalachiana* Triplett et al., were selected. Finally, from one lineage of the deeply diverging grade of grasses *Pharus latifolius* and the congeneric species *P. lappulaceus* Aubl. were selected. By assessing the accuracy, completeness, and efficiency of these assembly methods we compared different approaches to assembly and established guidelines for full plastome determination of grasses using short reads of approximately 100 base pairs (bp) produced from single-read libraries.

METHODS

DNA extraction and Sanger sequencing.

Silica dried leaf samples were obtained from five species: *Arundinaria appalachiana*, U.S.A., J. Triplett JT099 (ISC); *A. tecta*, U.S.A., J. Triplett JT173 (ISC); *Neyraudia reynaudiana*, China, J.T. Columbus 5302 (RSA); *Pharus latifolius*, U.S.A., J. Triplett 421 (MO); and *P. lappulaceus*, U.S.A., J. Triplett 422 (MO). Leaf tissue was homogenized manually in liquid nitrogen before extraction. The DNA extraction protocol using the Qiagen DNeasy Plant Mini Kit (Qiagen, Valencia, California, U.S.A.) was followed. The extraction from fresh leaf tissue of a sixth species, *Arundinaria gigantea* was performed as described in Burke et al. (2012).

Methods for obtaining complete plastomes using Sanger sequencing generally followed those described by Dhingra & Folta (2005). Overlapping segments of these plastomes were amplified using universal plastome primers for the inverted repeat (IR) regions (Dhingra & Folta, 2005) and primers specific for other regions of the grass plastome (Leseberg & Duvall, 2009). Each primer pair flanks a region of approximately 1200 bp. There are 125 such regions, 28 of which lie within the major inverted repeat and do not require duplicate sequencing except to locate the IR boundaries. Touchdown PCR was performed with all primer pairs using the "round I" conditions described by Dhingra & Folta (2005). Failure of PCR when using the main set of primers was addressed with the alternative methods of Morris & Duvall (2010) including the design of species-specific primers (Burke et al., 2012). Amplicons were purified using the Wizard SV PCR Clean-up System (Promega, Madison, Wisconsin, U.S.A.) and sent for Sanger sequencing at Macrogen (Seoul, South Korea). Quality of sequence information was verified and sequence identities were confirmed in duplicated bidirectional and overlapping sequences. Sequence assembly was performed using Geneious Pro v.6.1.6 (Biomatters, Auckland, New Zealand). All manual sequence manipulations in this study were performed using the Geneious Pro software package. A draft plastome of *Neyraudia reynaudiana* (88% complete) was determined and complete plastomes were determined for *Arundinaria gigantea* (Burke et al., 2012) and *Pharus latifolius* (Jones et al., 2014).

Next-generation sequencing (NGS).

Starting quantities of total genomic DNA from *Neyraudia reynaudiana*, *Arundinaria appalachiana*, *A. tecta*, and *Pharus lappulaceus* were determined by measurement at A260 with a Nanodrop 1000 (ThermoFisher Scientific, Wilmington, Delaware, U.S.A.) and diluted to contain approximately 1.5 µg each. DNA was diluted to approximately 2 ng/µl and sheared into ~300 bp fragments using a Bioruptor sonicator (Diagenode, Denville, New Jersey, U.S.A.) in two 12 min periods, inverting the tubes between periods. Sonicated DNA preparations were purified and concentrated with the MinElute Extraction Kit (Qiagen). Single-read libraries were prepared using the TruSeq sample preparation low-throughput protocol (gel method) following manufacturer instructions (Illumina, San Diego, California, U.S.A.). Sequencing was performed on a HiSeq 2000 instrument (Illumina) at Iowa State University (Ames, U.S.A.). Reads produced by this method were 99 bp in length. The single-reads were first quality filtered using DynamicTrim v.2.1 from the SolexaQA software package (Cox et al., 2010) with default settings, and then sequences shorter than 25 bp in length (default) were removed with LengthSort v.2.1 from the same package. The quality of the reads was then assessed using FastQC v.0.10.1 (http://www.bioinformatics.babraham .ac.uk/projects/fastqc/).

De novo assembly of NGS reads into contigs

(1) Greedy assembly. — The de novo assembly program in the Geneious Pro v.6.1.6 package provides a greedy assembly algorithm, similar to that of a multiple sequence alignment, which aligns all possible combinations of reads and assesses identity to produce

contigs (http://www.geneious .com). Because of the computational burden associated with greedy assembly on complete sets of reads, further filtration was performed on each set of reads to include only those with high identity to published plastome sequences. This filtration step was performed using the BLASTn software package (Altschul et al., 1997), which used a full plastome as a query against the read file database and an e-value threshold of 10^{-3}. The full plastome for *Panicum virgatum* L. (NC015990) was used as a query for the *N. reynaudiana* reads, *Bambusa oldhamii* Munro (NC012927) was used as a query for the *A. tecta* and *A. appalachiana* reads, and *Anomochloa marantoidea* Brongn. (NC014062) was used as a query for the *P. lappulaceus* reads. These taxa were used as a reference instead of the available conspecific or congeneric plastomes to simulate a more likely situation in which a plastome from the same species or congener would not typically be available. No filtration step was used prior to the other de novo assembly methods. Python scripts were used to extract matched sequences from the read files. The greedy assembly software was used to assemble these extracted sequences into contigs.

(2 and 3) Single-pass de novo assembly using de Bruijn-based assemblers. — There are a large number of de Bruijn-based assembly packages and we screened four of these in preliminary assemblies including Edena v.3 (Hernandez et al., 2008), SOAPdenovo v.1.02 (Li et al., 2010), SPAdes v.2.4.0 (Bankevich et al., 2012; http://bioinf.spbau.ru/spades), and Velvet v.1.2.08 (Zerbino & Birney, 2008; http://www.ebi.ac.uk/~zerbino/ velvet/). Two packages, SPAdes (2) and Velvet (3), which were specifically designed for the assembly of small genomes, were ultimately selected based on larger contig sets that had high plastome homology and speed of assembly,

although some other assembly packages may also be appropriate for use in plastome determination.

For purposes of comparison, the parameters for both programs were set to twelve identical k-mer lengths (units of read overlap). A minimum k-mer length of 19 was chosen and increased by steps of 6 bp 11 times until the maximum k-mer length, 85, was reached. All k-mer lengths were used and merged in the same assembly run. The shorter k-mer length is used to achieve assembly in areas of lower coverage while the longer k-mer length is used to keep this method consistent with the next section in which longer contigs as used as input. Although this procedure includes 12 assemblies, we refer to it as a "single pass" assembly, to distinguish it from our alternative approach ("iterative assembly", see below). One Velvet assembly was performed for each k-mer length and the contigs generated from each of these were pooled into single files. Note that SPAdes performed this task automatically. The N50 of each contig set was calculated using a Python script.

Iterative de novo assembly. — The contigs generated by performing single-pass de novo assemblies using Velvet were assembled into larger contigs using the same procedure. This method could not be performed with the SPAdes software package because of a software limitation on user-provided sequence length, in which initially assembled contig files could not be reassembled. Assemblies were repeated until substantially larger contigs were no longer generated, which occurred after the third assembly. The output from each step was pooled into one file and assembled de novo one additional time. These results were then pooled with the previous de novo assembly results because sequences that are not incorporated into Velvet assemblies are excluded from the output files. To summarize, in this

iterative approach, assemblies were performed three times and were finally applied to all intermediate de novo results in a fourth and final Velvet assembly (Fig. 1). Python scripts were used to automate this repetitive task.

Contig scaffolding

(4) Map to reference. —One IR region was omitted from the reference plastome to reduce the number of erroneous hits to the inverted repeat boundaries. The Geneious Pro software package was used to map the contigs generated onto their positions within plastomes. Each contig set produced by de novo assembly was mapped to its respective full plastome reference that was used to query the reads file using BLASTn as described above. While the mapping methods of Steele et al. (2012), which used the published plastome from each taxon as a reference, would allow for more accurate assemblies, we used closely related reference plastomes instead for reasons stated above.

(5) In silico genome walking. — Contigs generated using de novo assembly were scaffolded by locating overlapping regions. First the longest contig with plastome homology within Poaceae was selected. A region of 20–45 bp in length at the end of each contig was used as a query in the pool of generated contigs. When an exact match was located, the contig that contained this match was concatenated to the end of the initial contig (minus the overlapping region). Genome walking was performed both manually and using Python scripts to automate the process.

(6) Anchored conserved region extension (ACRE). — A rough-draft plastome alignment of 75 taxa within Poaceae was generated (Duvall et al., unpub.) using MAFFT

v.1.2 (Katoh et al., 2005). MAFFT was also used for all subsequent alignments in this study. Regions that were identically conserved among all 75 taxa and were greater than 19 bp in length were put into an input file. This length was chosen because it is long enough to reasonably assess homologous regions and short enough to locate complete family-specific conserved regions. For each of these 85 regions, the largest contig that included the region was located and scaffolded in the order in which they appeared in the plastome by combining overlapping regions. This original ACRE method was then automated as a Python script. These scripts as well as the ones used for in silico genome walking can be found at http://sourceforge .net/projects/grassplastome/.

Final assessment

Final plastome assembly. — The quality of all preliminary scaffolds was assessed by aligning the scaffolds to the respective reference plastome used previously for read filtration using MAFFT. This allowed for the percentage of each plastome assembled to be calculated for each contig assembly method. Each plastome assembly that covered more than 80% of its reference was used in subsequent analysis. Missing regions in each assembly were inserted manually by locating flanking overlapping regions from each respective read file or the combined contigs file for larger gaps. Large changes such as indel mutations greater than 20 bp in length or large regions of substitutions were verified by locating overlapping sequences in the original read file. Final plastome assemblies were arranged and trimmed to a proper orientation for sequence alignment. The 5′ end of the large single-copy region was positioned on the 5′ end of the total assembly while the 5′ end of the inverted repeat region A (IRa) was

positioned at the 3′ end of the total assembly. The IRa region was omitted from alignment because the inclusion of two inverted repeats would double the representation of that region during subsequent analyses. These boundaries were located using the methods in Burke et al. (2012) by identifying the region where the end of the sequence matches the 3′ IRb boundary and where the beginning of the sequence meets the 5′ IRb boundary using BLAST.

Assessment of contig assembly methods. — Assemblies generated using each method were assessed using one criterion for accuracy, the percent identity that the assembly shared with its partnered Sanger-sequenced plastome, and two criteria for completeness, the quantity of missing sequence in the plastome prior to final plastome assembly and the number of ambiguous nucleotide sites that could not be resolved by the contig assembly. Percent identity was assessed by aligning the completed plastome generated using NGS data to either the Sanger-sequenced plastome of the same species (*N. reynaudiana*) or the Sanger-sequenced plastome of its congener (*A. appalachiana*, *A. tecta*, *P. lappulaceus*). Nucleotide sites in the alignment of the *N. reynaudiana* plastomes, where gaps were present in the Sanger sequence, were omitted so that only the Sanger-sequenced 88% of the plastome was represented while calculating the percent similarity.

Assessment of read depth. — As a final assessment, each complete plastome was subjected to a reference mapping using only each respective trimmed and filtered read set. Because of the unusually low number of reads associated with *A. tecta*, the mapping for this species used a more sensitive setting along with a gap optimization step. This allowed for the mean, minimum and maximum read depth to be calculated using the Geneious Pro software. By mapping the reads onto the final assembly we ensured that read depth measurements were

accurate. A consensus sequence was also generated from the read mapping, which was aligned to the assembled plastome as further verification.

RESULTS

Initial sequencing results

Sequences of complete plastomes were determined and deposited in GenBank for five species—*Neyraudia reynaudiana*, *Arundinaria appalachiana, A. tecta*, *Pharus latifolius*, and *P. lappulaceus*. Note that the sixth plastome from *Arundinaria gigantea* was previously sequenced and assembled (Burke et al., 2012). GenBank accession numbers and comparative lengths of plastome regions and subregions are given in Table 1.

After trimming for quality and filtering out short reads, the single-read files for *N. reynaudiana*, *A. appalachiana*, *A. tecta*, and *P. lappulaceus* contained 7.96, 5.42, 0.92 and 1.75 million reads, respectively. Fragments in each set were 25–99 bp in length, had a median length of 99 bp, and a mean length of 93–94 bp. After de novo assembly (Methods 1–3; Fig. 2) the longest contig in each set with grass chloroplast homology ranged from 2699 to 48,146 bp in length. The N50 of each contig set ranged from 86 to 8114. The results for both of these measurements are reported in Table 2.

Out of the 48 combinations of de novo assembly and contig scaffolding methods tested here, 12 of these met the 80%plastome coverage threshold. Note that the contigs generated by the de novo greedy assembly method were the shortest, had the most missing and ambiguous sites and produced the lowest percent plastome coverage after scaffolding.

Because the percent plastome coverage did not exceed the target 80%, full plastomes were not generated using these contig sets.

Assessment of contig scaffolding methods.

(4) Map to reference. — This method generated entire plastome sequences that possessed similarity with each corresponding Sanger-sequenced plastome ranging from 39.98% to 98.87% before manual attention was given to the sequence and a final assembly was produced. These sequences were missing or ambiguous at 297–56,413 nucleotide sites with the greatest number of these sites being generated when SPAdes was used to assemble the reads de novo even though this method often produced the longest contigs. The results for each map to reference assembly are reported in Table 3. The mappings that were manually resolved into a final assembly after using Velvet iteratively for *N. reynaudiana*, *A. appalachiana*, *A. tecta*, and *P. lappulaceus* showed 99.38%, 99.06%, 98.96%, and 98.94% similarity (Table 4), respectively, to their partnered Sanger-sequenced plastome. These were the only set of reference mappings that were assembled into full plastomes because Velvet produced the highest amount of coverage for each species (Table 3) and the iterative strategy produced larger contigs than the single-pass strategy (Table 2).

(5) In silico genome walking. — This method initially generated greater than 80% of a plastome for only one set of contigs (*N. reynaudiana* assembled de novo using Velvet iteratively) and substantially less for the rest of the contig sets (Table 5). The scaffolded contigs from *Neyraudia* assembled using Velvet also required regions to be manually rearranged to conserve gene order and complete both inverted repeat sequences. After

rearrangement of regions and final assembly was performed, this method produced an assembly that was 99.46% similar to its partnered partially Sanger-sequenced plastome.

(6) ACRE. — This assembly method was attempted with the contigs generated using both Velvet and SPAdes for each of the four taxa. Before any manual attention was given to produce a final assembly, these methods generated from 99.63% (*N. reynaudiana* contigs assembled de novo using Velvet iteratively) to 4.11% (*A. tecta* contigs assembled de novo using SPAdes) of the full plastome. Table 5 reports these percentages. When the assemblies from *A. tecta* were omitted from the set, using the ACRE method yielded 81.08%–99.63% of the plastome automatically. Because this analysis concatenated overlapping contigs with matching motifs, it also generated no ambiguous nucleotide sites. After final assembly was completed, ACRE methods produced assemblies that were 98.48% to 99.66% similar to their partnered Sanger-sequenced plastomes (Table 4).

Read depth. — Read depth varied between taxa, but stayed consistent within taxa for each of the different assembly methods. Mean read depth values fell within a broad range from 16.4 to 132.7. Mean, minimum, and maximum read depth for each assembled plastome are reported in Table 4.

DISCUSSION

Methods to assemble short NGS reads economically, accurately, rapidly, and in a largely automated manner with little subsequent need for manual adjustment are key to the successful use of plastomes for grass systematics. The use of single-read libraries can economize the production of next-generation sequence files. Over the past eight years, many

de Bruijn based de novo assembly software packages have been released (Zhang, W. et al., 2011) and have been shown to assemble short reads effectively and conservatively. However, as the reads are assembled conservatively into contigs, complete plastomes are not always produced, especially when single-read technology, which is more cost effective under certain conditions, is used (as opposed to paired-end reads). This creates a need to assemble them using somewhat less stringent methods. The contigs assembled by this type of software can be used to assemble a smaller genome by aligning them to a closely related taxon where gene order is largely conserved, as in most Poaceae, and by filling in any gaps with reads or smaller contigs using flanking overlap. While this method is fairly simple and accurate it can be very time consuming and laborious when applied to a larger-scale study and may, in some cases, introduce bias.

The assemblies produced using each method were assessed using three criteria: 1) their identity with each partnered Sanger plastome; 2) the quantity of missing sequence; and 3) the number of ambiguous sites. Because of the longer reads and targeted nature of Sanger sequencing, a plastome sequenced using this method can be a reliable aid in testing whether shorter and randomly targeted NGS reads are assembled accurately. Since an accurate assembly is crucial in subsequent phylogenomic analyses, percent similarity to its partnered Sanger sequence is the most important criterion. The efficiency of the assembly, which can be quantified using the number of gaps and ambiguous sites, is of less concern because complications can be easily identified and resolved. However, this does present a concern because of the amount of time and labor required to accurately perform these adjustments. At first glance, repairing a small number of assemblies seems manageable, but as the number of

assemblies grows, this task becomes an increasing hindrance to achieving final downstream analyses.

One criterion that was not emphasized as an assessment of assembly quality here is the N50 of each contig file. The N50, used in many NGS assembly studies, is a weighted median statistic for assessing the distribution of contig lengths, where a higher value reflects a greater proportion of longer contigs. Studies with the objective of generating the largest possible contigs for purposes of assembling large sequences such as eukaryotic chromosomes (e.g., Li et al., 2010; Nowrousian et al., 2010) require de novo assembly of a contig set with a higher N50. In the study presented here, large contigs are useful in establishing preliminary assemblies while the smaller contigs are also useful for gap bridging and resolution of ambiguities. In addition, the N50 statistic (Table 2) is not comparable between different assemblies if their combined lengths are not equal (Miller et al., 2010). Other factors such as the presence of nuclear, mitochondrial, and microbial sequences within each contig set also make the comparison of N50 values between assembled contig sets less useful for the methods of grass plastome assembly considered here, since a large contig could potentially represent one of these other sources of DNA.

After completion, *N. reynaudiana* and *A. appalachiana* exhibited greater than 99% similarity with the Sanger-sequenced plastome. Plastome assemblies for *A. tecta*, and *P. lappulaceus* exhibited greater than 98% similarity with each of their Sanger-sequenced congeneric plastomes (Table 4). While two plastomes from the same individual would be expected to be indistinguishable regardless of which sequencing method was used, sequencing artifacts such as erroneous base calls by capillary sequencing software and

incorrectly incorporated bases during early stages of PCR can result in low frequency nucleotide polymorphisms between the two assemblies. The partially Sanger-sequenced plastome for *N. reynaudiana* and its next-gen assemblies contained polymorphisms at less than 0.5% of the nucleotide sites.

Prior to contig scaffolding, methods for de novo assembly of reads exhibited no clear patterns in effectiveness. The plastome for *N. reynaudiana* was most accurately assembled using Velvet iteratively, but was only marginally more accurate than the single-pass strategy using either of the de Bruijn assemblers used here. Although using single-pass Velvet generally allowed for fewer ambiguities after subsequent reference mapping (Method 4), this strategy did not produce more than 56% of a complete plastome when combined with genome walking (Method 5) or ACRE (Method 6) in the other three taxa. The SPAdes software package did perform somewhat more accurately and with comparable efficiency than the iterative Velvet approach when combined with the ACRE method in *A.appalachiana* and *P. lappulaceus*. However, the iterative Velvet approach executed more rapidly than SPAdes assemblies.

Performing the ACRE method after running Velvet iteratively or SPAdes single-pass also successfully revealed a 596 bp insertion in the *psbE-petL* intergenic spacer of *A. appalachiana*. This sequence was absent in the Sanger-sequenced *A. gigantea*, but found in all other Arundinarieae (Burke et al., 2012). Note that a reference-guided assembly of the plastome of *A. appalachiana* using *A. gigantea* as a reference failed to detect this large insertion because of the bias imposed by the reference (Wysocki, unpub. obs.). This

shortcoming of reference mapping assembly demonstrates that improvements can be made to assemblies by using methods that rely less on reference sequence data.

The reference mapping scaffolding method did generate substantial amounts of the plastome sequence, however the sheer number of ambiguous nucleotide sites presents a problem for efficient assembly. To generate a full and accurate plastome, each of these ambiguous sites would require a verification using sequences within the original read file. This would require the arduous, time-consuming and computationally taxing endeavor of querying millions of reads for motifs flanking each of the potentially thousands of ambiguous sites. Note that reference assembly cannot be accomplished if no closely related reference sequences are available.

Although the reference mapping method (Method 4) can be labor intensive, it does occasionally possess an indispensable role in plastome assembly. The contigs generated from the *A. tecta* reads did not contain enough coverage or overlap to complete a substantial amount of the plastome sequence using genome walking or ACRE, likely because the original read file contained the fewest reads. Reference mapping after using the iterative Velvet strategy did produce a substantial amount of the plastome and by manually repairing the ambiguities a full plastome was able to be generated.

In silico genome walking (Method 5) can function well as a method for de novo assembly if contigs are well represented around the entire plastome. However, genome walking will fail in areas that contain little overlap, even if larger areas are well represented. The overlap does require an exact sequence match, which presents a problem for regions that contain ambiguities in poorly covered areas of the plastome. This method also exhibits

problems when approaching IR region boundaries. Genome walking software cannot distinguish between assembling a plastome downstream in one IR region and upstream in the other copy. Even when the repeat regions are assembled correctly, a product of this type of assembly performed on a circular sequence may also produce a fragment with regions in an order that may need to be manually rearranged for alignment purposes. While these complications can be remedied by locating the boundaries of each region of the plastome and disassembling the sequence and placing each region where it belongs, this also consumes time and labor and cannot be conventionally streamlined. One benefit of genome walking is that preliminary data, such as draft or reference plastomes, are not necessary. Plastome assembly using genome walking only requires that a plastome-homologous sequence that was assembled de novo from reads is identified. This method could potentially allow for de novo assembly of plastomes within taxa that lack a Sanger-sequenced reference plastome and in which gene order may not be preserved, such as between the plastomes of Poaceae and non-grass monocots. Another application for genome walking is its ability to span highly variable regions where a reference sequence could not be utilized. This could allow for nuclear genes with large introns to be extracted from contig assemblies and the reads themselves.

While most complete assemblies performed here showed similarity greater than 98%, the ACRE method (6) produced full plastome assemblies that required the least amount of manual attention. The ACRE method can be effective in rapidly generating accurate datasets for large-scale intrafamilial analyses when preliminary data from previously sequenced plastomes that allow identification of conserved regions are available, as from Poaceae,

Asteraceae, and Fabaceae. The abundance of preliminary data for these families can be attributed in part to their large size and widespread scientific interest. However, this type of assembly will become applicable to other families as more plastomes are sequenced and conserved regions can be identified. The utility of this type of assembly can also be attributed to the map to reference method. While a quick map to reference on many taxa without manual resolution of ambiguous sites can produce erroneous results in phylogenomic analyses and will not clarify some of the unique features of a sequence, it can aid in finding ultra-conserved regions within a small genome. These regions are fundamental to performing the ACRE assembly method, which can effectively compensate for the previously mentioned weaknesses of a map to reference.

CONCLUSIONS

When performing a phylogenomic study that includes numerous taxa, automation of assembly becomes essential to eliminate tedious and time-consuming steps. Because of this, small genomes should be assembled first using the most automatable methods (ACRE, in silico genome walking), which can largely complete the assembly using read files with high coverage and overlap. This can be followed by less automatable processes (map to reference, manual assembly), which may be necessary for taxa that possess a smaller number of reads and do not produce successful results when subjected to automatable processes (Fig. 2).

For purposes of measurement, methods performed in this study were kept consistent and comparable. However, a variety of adjustments could be applied to the protocols outlined here to ensure that more successful, efficient, and complete assemblies can be performed.

One obvious adjustment is the insertion of all of the reads into each respective contig file prior to assemblies such as mapping to a reference or in silico genome walking. These reads may ensure that overlapping regions with insufficient read depth, which may have been lost during initial de novo assembly, are represented and can potentially eliminate ambiguities and holes. Another potential adjustment is to combine two or more contig sets that were generated from the same reads file, but used more than one de novo assembly software package or algorithm.

Phylogenomic studies are quickly becoming more large-scale and widespread. This makes efficient and high-throughput pipelines for assembling short-read sequences increasingly crucial. In Poaceae, applications range from agrostology, biodiversity and bioconservation studies, bioengineering and functional ecology. While the scope of the aforementioned methods may be limited to highly conserved plastomes, they can be utilized to perform other tasks by altering parameters. As technological capabilities increase and allow for longer sequencing reads and more advanced computing power, modified versions of the methods used in this paper can be put to use for plastome determination with more efficient assemblies that produce more accurate results.

Table 1. GenBank accession numbers and lengths of regions and subregions for plastomes used in this study.

Lengths (bp):		**Total**	**LSC[a]**	**SSC[b]**	**IR[c]**
Neyraudia reynaudiana	KF356392	135,367	80,616	12,695	21,028
Arundinaria gigantea	JX235347	138,935	82,641	12,700	21,797
A. appalachiana	KC817462	139,547	83,223	12,717	21,804
A. tecta	KC817463	139,499	83,162	12,730	21,804
Pharus latifolius	JN032131	142,077	83,341	12,530	23,103
P. lappulaceus	KC311467	141,928	83,188	12,536	23,102

[a]Large Single-Copy Region
[b]Short Single-Copy Region
[c]Inverted-Repeat Region

Table 2. The N50 values and largest contigs generated for each set of reads using the Greedy assembly function included in the Geneious Pro package, SPAdes, and Velvet. Velvet assemblies were performed using both the single-pass and iterative methods.

Species	***De Novo Assembly Method***	***N50***	***Largest Contig (bp)***
Neyraudia reynaudiana	Geneious	388	3431
	SPAdes	401	48146
	Single-Pass Velvet	125	24265
	Iterative Velvet	153	36174
Arundinaria appalachiana	Geneious	8114	28388
	SPAdes	506	28504
	Single-Pass Velvet	86	20022
	Iterative Velvet	105	23075
A. tecta	Geneious	281	2823
	SPAdes	506	4817
	Single-Pass Velvet	93	2699
	Iterative Velvet	99	4991
Pharus lappulaceus	Geneious	1327	8554
	SPAdes	395	18863
	Single-Pass Velvet	97	5429
	Iterative Velvet	106	13154

Table 3. Summary of the results of the contig scaffoldings performed using the reference mapping function in Geneious Pro v.6.1.6 after de novo assembly using Velvet, SPAdes, or the assembly function included in the Geneious Pro package. Velvet assemblies were performed using both the single-pass and iterative methods.

Species	***De Novo Assembly Software***	***Ambiguous Nucleotide Sites***	***% Identity w/ Sanger-Sequence***
Neyraudia reynaudiana[a]	Iterative Velvet	1898	97.01
	Single-Pass Velvet	577	98.87
	SPAdes	47486	53.74
	Geneious	44455	65.72
Arundinaria appalachiana	Iterative Velvet	355	97.61
	Single-Pass Velvet	297	97.40
	SPAdes	18985	81.34
	Geneious	16371	85.33
A. tecta	Iterative Velvet	22228	70.04
	Single-Pass Velvet	34689	69.29
	SPAdes	56413	39.98
	Geneious	31524	72.21
Pharus lappulaceus	Iterative Velvet	1469	92.11
	Single-Pass Velvet	3114	91.07
	SPAdes	11566	84.84
	Geneious	15647	82.95

[a] Percent similarity with the Sanger plastome for *Neyraudia* was calculated over the 88% completed.

Table 4. The similarity between plastome assemblies generated using NGS data and their partnered Sanger-sequenced plastome. Each assembly method includes a de novo assembly followed by contig scaffolding. Plastome assemblies that were less than 80% complete after contig scaffolding are not included here.

Species	***Assembly***	***% Identity with Sanger***	***Polymorphic Nucleotide Sites***	***Read Depth (per bp)***		
				Mean	**Min**	**Max**
Neyraudia reynaudiana[a]	Velvet (iter)[b] - ACRE	99.66	343	132.7	33	439
	Velvet (SP)[c] - ACRE	99.66	345	132.5	1	439
	SPAdes - ACRE	99.66	344	132.7	33	439
	Velvet(iter)-Walking	99.66	549	132.7	33	439
	Velvet (iter)-MTR[d]	99.38	618	129.8	2	445
Arundinaria appalachiana	Velvet (iter)-ACRE	99.24	889	16.8	1	1725
	SPAdes-ACRE	99.35	766	16.4	1	171
	Velvet (iter)-MTR[b]	99.06	1102	16.4	1	170
A. tecta	Velvet (iter)-MTR	98.96	1228	31.8	1	709
Pharus lappulaceus	Velvet (iter)-ACRE	98.48	1542	18.8	1	86
	SPAdes - ACRE	98.54	1739	18.8	1	86
	Velvet (iter)-MTR	98.94	1259	18.9	1	86

[a] *N. reynaudiana* % identity with its partnered Sanger-sequenced plastome was calculated using the 88% of the existing, Sanger-sequenced plastome. [b] Velvet(iter): Velvet run iteratively according to methods outlined in Figure 1. [c] Velvet(SP): Velvet run once.

[d] MTR: Map to reference method

Table 5. Percent of each Sanger-sequenced plastome covered by first de novo assembling using Velvet, SPAdes, or the greedy assembly function included in the Geneious Pro package on each set of reads and then applying the ACRE or *in silico* genome walking scaffolding method to each set of contigs. Velvet assemblies were performed using both the single-pass and iterative methods.

Species	***De novo* assembly method**	**ACRE (%)**	**Walking (%)**
Neyraudia reynaudiana	Iterative Velvet	99.63	89.62
	Single-Pass Velvet	95.85	24.52
	SPAdes	98.75	49.72
	Geneious	25.12	2.92
Arundinaria appalachiana	Iterative Velvet	92.94	23
	Single-Pass Velvet	55.19	20.29
	SPAdes	81.08	23.93
	Geneious	79.29	24.18
A. tecta	Iterative Velvet	7.79	4.26
	Single-Pass Velvet	14.87	2.64
	SPAdes	17.36	4.11
	Geneious	18.54	2.4
Pharus lappulaceus	Iterative Velvet	84.12	13.77
	Single-Pass Velvet	31.83	8.43
	SPAdes	93.6	18.19
	Geneious	62.99	7.16

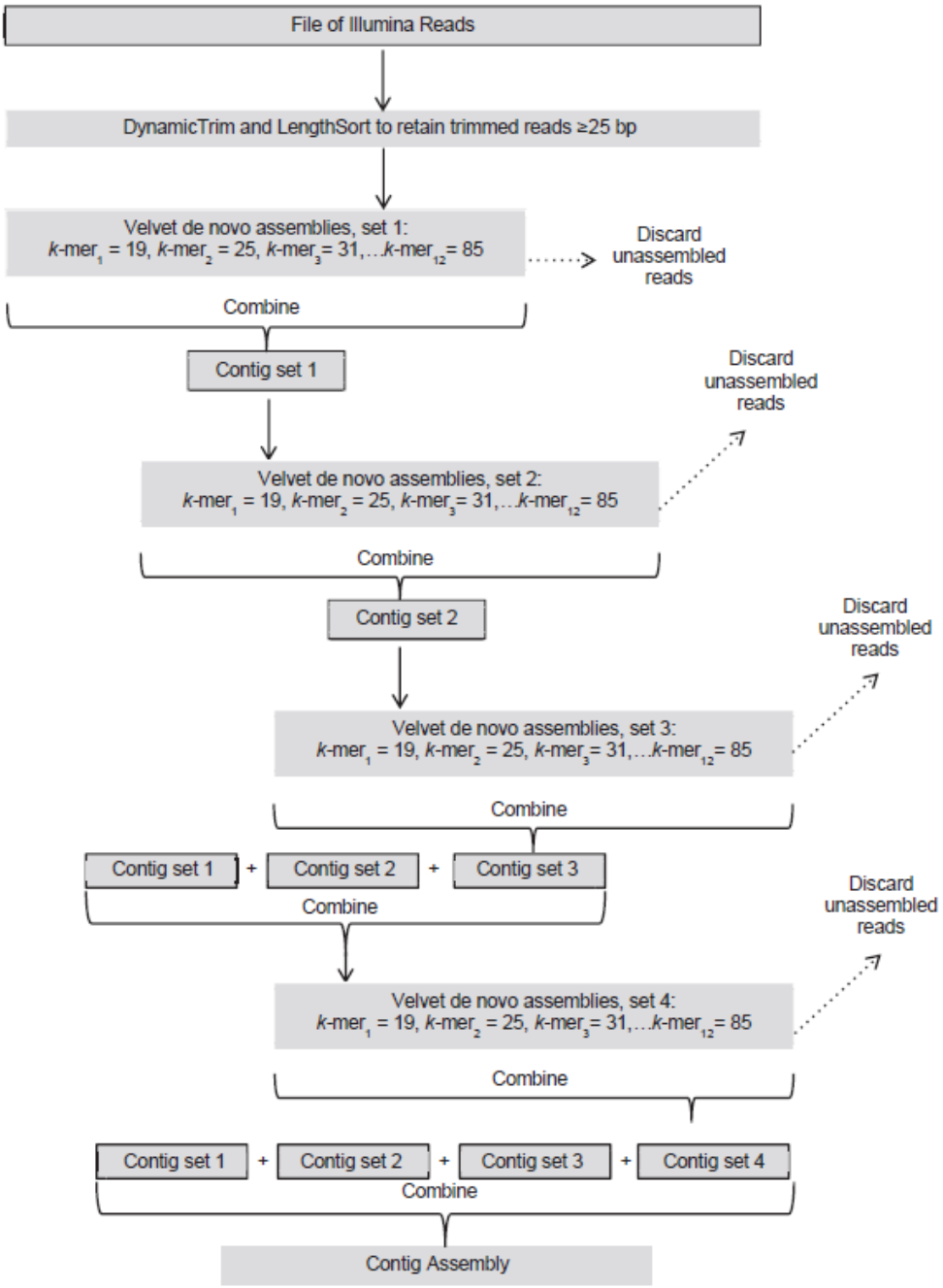


Figure 1. Summary of iterative de novo assembly using Velvet. Boxes with borders indicate either read or contig files. Boxes without borders indicate processes. Points at which contig files are combined are indicated.

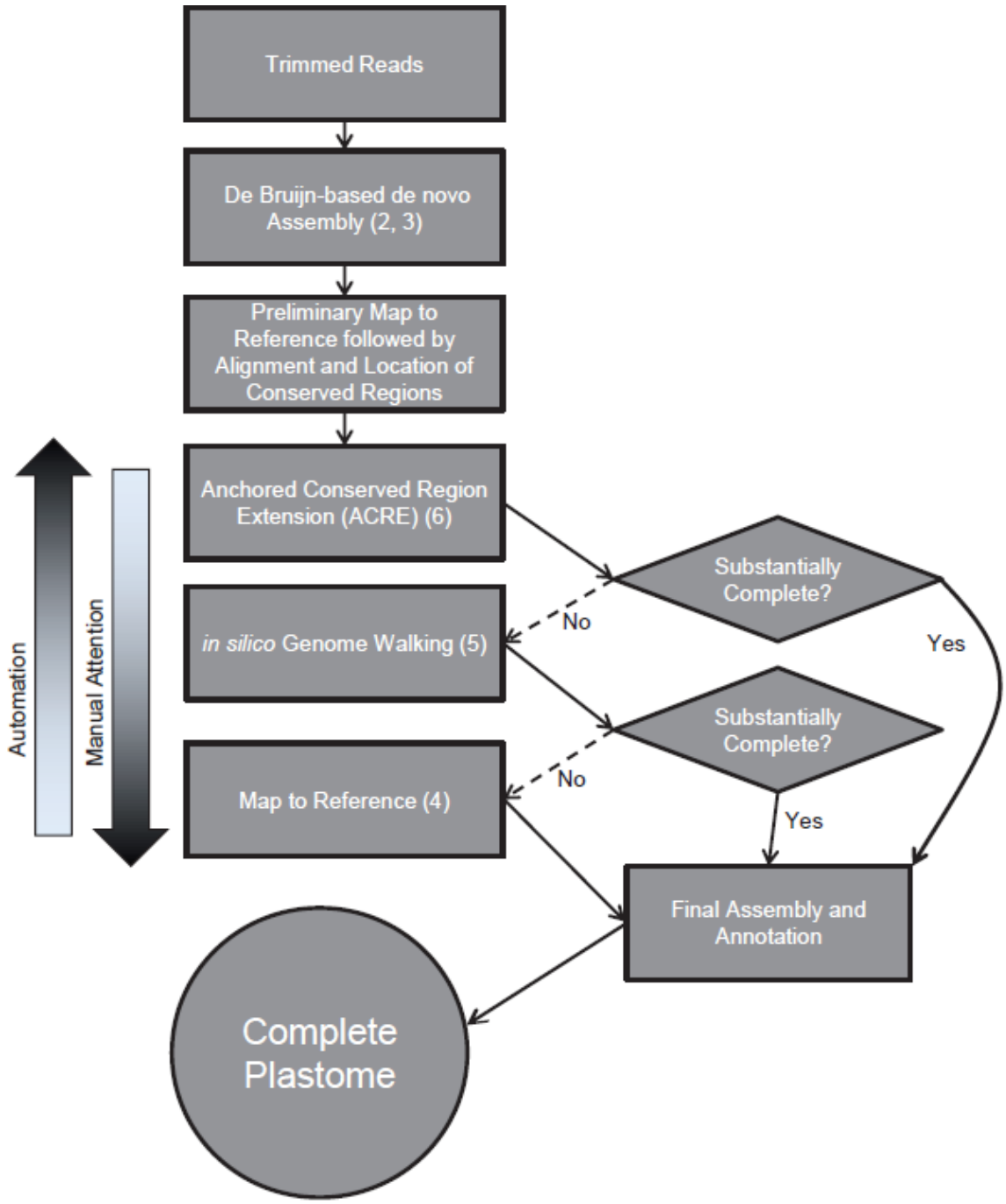


Figure 2. Overall summary of short read plastome assembly process. Numbers in each box correspond to each assembly tool or method as outlined in the Materials and Methods section. Note that method 1, greedy assembly, does not appear in this figure. Shaded arrow directionality indicates increasing automation or required manual attention.

CHAPTER 2

EVOLUTION OF THE BAMBOOS (BAMBUSOIDEAE; POACEAE): A FULL PLASTOME PHYLOGENOMIC ANALYSIS

ABSTRACT

Bambusoideae (Poaceae) comprise three distinct and well-supported lineages: tropical woody bamboos (Bambuseae), temperate woody bamboos (Arundinarieae) and herbaceous bamboos (Olyreae). Phylogenetic studies using chloroplast markers have generally supported a sister relationship between Bambuseae and Olyreae. This suggests either at least two origins of the woody bamboo syndrome in this subfamily or its loss in Olyreae. Here a full chloroplast genome (plastome) phylogenomic study is presented using the coding and noncoding regions of 13 complete plastomes from the Bambuseae, eight from Olyreae and 10 from Arundinarieae. Trees generated using full plastome sequences support the previously recovered monophyletic relationship between Bambuseae and Olyreae. In addition to these relationships, several unique plastome features are uncovered including the first mitogenome-to-plastome horizontal gene transfer observed in monocots. Phylogenomic agreement with previous published phylogenies reinforces the validity of these studies. Additionally, this study presents the first published plastomes from Neotropical woody bamboos and the first full plastome phylogenomic study performed within the herbaceous bamboos. Although the phylogenomic tree presented in this study is largely robust, additional studies using nuclear genes support monophyly in woody bamboos as well as hybridization

among previous woody bamboo lineages. The evolutionary history of the Bambusoideae could be further clarified using transcriptomic techniques to increase sampling among nuclear orthologues and investigate the molecular genetics underlying the development of woody and floral tissues.

INTRODUCTION

Bambusoideae are a lineage of perennial forest grasses (Poaceae) endemic to every continent except Europe and Antarctica (Bamboo Phylogeny Groupp (BPG), 2012, Kelchner and BPG, 2013). The Bambusoideae comprise 115 genera and approximately 1450 species of bamboos (BPG, 2012). Bambusoideae are divided into two morphologically distinct habits: woody and herbaceous bamboos. While woody bamboos display a wide range of morphological diversity, they do possess multiple shared characteristics. The woody bamboo syndrome includes strongly lignified culms, specialized culm leaves, complex vegetative branching, outer ligules on the foliage leaves, bisexual flowers, and gregarious monocarpy (BPG, 2012). These bamboos, some of which can quickly grow up to 45 m in height, serve as an economically important source of building materials and other products for cultures in Central and South America, Asia, Africa, and Australia (http://www.eeob.iastate.edu/research/bamboo/bamboo.html). Their potential for rapid establishment combined with their extensive vegetative reproduction also make bamboos important ecologically as they can serve as forest habitats of their own and can affect the survival of sympatric woody species (Lima et al., 2012). The gregarious, semelparous flowering patterns of woody bamboos and subsequent death of the parent plant can have

ecological effects such as the increase in pest populations during the fruiting of *Melocanna baccifera* in regions of India (Sarma, 2009) and the increase in eudicot sapling growth during the die-off of the dominant forest bamboo *Chusquea culeou* (Marchesini et al., 2009). This pattern of flowering is correlated with increased generation times in this group, which in turn is correlated in bamboos and other grasses with shorter branch lengths in phylogenetic analyses (Gaut et al., 1997) and fewer resolved nodes between certain closely related species.

Herbaceous bamboos are characterized by shorter and more weakly lignified shoots, less vegetative branching, unisexual flowers, and annual or seasonal flowering patterns (BPG, 2012). The flowering phenology of herbaceous bamboos is correlated with an increase in the substitution rates observed in chloroplast loci. This has at least two consequences relevant to bambusoid plastome phylogenomics. First, phylogenetic resolution and support within this group are likely to be increased due to higher numbers of informative sites. At the same time, long branches are produced with the potential for long-branch attraction artifacts between herbaceous bamboos and non-bambusoid outgroups. Phylogenomic results can be more realistically interpreted taking these effects into account.

Molecular studies have placed Bambusoideae in the BEP (Bambusoideae, Ehrhartoideae, Pooideae) clade of Poaceae. A sister group relationship between Bambusoideae and Pooideae has been strongly supported (Bouchenak-Khelladi et al., 2008: Zhang et al., 2011), although morphological synapomorphies have yet to be found that unite these two subfamilies. Bambusoideae can be divided into three well-supported monophyletic tribes: the woody Arundinarieae and Bambuseae, and the herbaceous Olyreae (BPG, 2012).

The Bambuseae are native to tropical areas in both the Old and New World. This tribe comprises two clades that correspond to Old and New World species (Kelchner and BPG, 2013; Hisamoto et al., 2008). Phylogenetic studies that use plastid markers generally place Olyreae as the sister group to Bambuseae in well-supported trees (Kelchner and BPG, 2013; Sungkaew et al., 2009; Burke et al., 2014). Olyreae are exclusively distributed in the New World except for *Buergersiochloa bambusoides*, a species endemic to New Guinea, and *Olyra latifolia*, which is found widely distributed in Africa/Madagascar as well as in the New World (Lovett, 1994). However, the African origin of the *O. latifolia* population has been debated (Soderstrom and Zuloaga, 1989). Like the Bambuseae, the Arundinarieae include woody bamboos found in both the Old and New World, with a basically Laurasian distribution pattern, but unlike most Bambuseae they are well-adapted to temperate environments.

Although paraphyly of the woody syndrome in Bambusoideae is well supported by tree analyses that use maternally inherited chloroplast phylogenetic markers (Bouchenak-Khelladi et al., 2008), this has been a subject of debate. Network analyses have revealed that the phylogenetic placement of Olyreae is less certain than previously reported (Kelchner and BPG, 2013).To be consistent with the chloroplast phylogeny, the woody bamboo syndrome would have either evolved twice independently (once in each of the ancestors of the Bambuseae and Arundinarieae) or arisen once in the common ancestor of the Bambusoideae and then subsequently been lost in the Olyreae. A hypothesized single origin of the woody bamboo syndrome, which has been most recently supported by Triplett et al. (2014), is evolutionarily more parsimonious than these scenarios.

In the past three years, full plastome phylogenomic analyses have been used to address evolutionary problems in the Bambusoideae. These analyses have been variously applied in Bambusoideae to resolve subfamilial relationships (Zhang et al., 2011; Wu et al., 2009; Wu and Ge, 2012) and investigate biogeographical patterns (Burke et al., 2012; Burke et al., 2014). Full plastome analysis can also provide enough information to resolve difficult interspecific relationships. This is an issue that is especially relevant to woody bamboos, which generally hybridize readily and exhibit very long generation times (McClure, 1966; Wong and Low, 2011). While studies such as Kelchner and BPG (2013) and Triplett & Clark (2010) have used selected chloroplast markers to infer maternally inherited evolutionary signal within Bambusoideae, our objective is to use all coding and non-coding regions within the chloroplast to increase the number of informative sites. Here, a full plastome phylogeny was generated using 13 tropical woody species, 10 temperate woody species and eight herbaceous species with 17 newly sequenced and 15 existing bambusoid plastomes plus two outgroup plastomes.

METHODS

Taxon sampling and DNA extraction

Silica-dried leaf tissue was obtained from 17 species of bamboos (*Bambusa arnhemica* F.Muell., *Bambusa bambos* (L.) Voss, *Buergersiochloa bambusoides* Pilg., *Chusquea liebmannii* E. Fourn. ex Hemsl., *Chusquea spectabilis* L.G. Clark., *Diandrolyra* Stapf. sp., *Eremitis* Döll sp., *Greslania* Balansa sp., *Guadua weberbaueri* Pilg., *Hickelia madagascariensis* A. Camus, *Lithachne pauciflora* P. Beauv., *Neololeba atra* (Lindl.) Widjaja, *Olmeca reflexa* Soderstr., *Otatea acuminata* (Munro) C.E. Calderón & Soderstr.,

Pariana radiciflora Sagot ex Döll, *Raddia brasiliensis* Bertol., and *Thamnocalamus spathiflorus* Munro) and one ehrhartoid species (*Zizania aquatica* L.). Herbarium voucher specimens were collected and are reported in Table 6. Tissue was homogenized manually in liquid nitrogen before extraction. The DNA extraction protocol using the Qiagen DNeasy Plant Mini Kit (Qiagen Inc., Valencia, CA) was followed.

Illumina sequencing and quality control

Starting quantities of total genomic DNA from *Bambusa arnhemica*, *B. bambos*, and *Thamnocalamus spathiflorus* were determined by measurement at A260 with a Nanodrop 1000 (ThermoFisher Scientific, Wilmington, DE, USA) and diluted to contain approximately 1.5 µg each. DNA was diluted to approximately 2 ng/µl and sheared into ~300 bp fragments using a Bioruptor® sonicator (Diagenode, Denville, NJ, USA) in two 12 min., periods, inverting the tubes between periods. Sonicated DNA preparations were purified and concentrated with the MinElute Gel Extraction Kit (Qiagen Inc., Valencia, CA, USA). Single read libraries were prepared using the TruSeq sample preparation low throughput protocol (gel method) following manufacturer instructions (Illumina, San Diego, CA, USA). Sequencing was performed on a HiSeq 2000 instrument (Illumina, San Diego, CA, USA) using single reads at the Iowa State University DNA Sequencing Facility, Ames, IA, USA. Reads produced by this method were 99 bp in length.

Quantities of total genomic DNA from *Chusquea liebmannii*, *Otatea acuminata* and *Pariana radiciflora* were determined using the Qubit fluorometric quantitation system (Life Technologies, Grand Island, NY, USA). Two micrograms were used in each library

preparation. Libraries were prepared using the TruSeq Nano DNA sample preparation kit (Illumina, San Diego, CA, USA) and sequenced paired-end at Cold Spring Harbor laboratory, Cold Springs, New York, USA.

Total genomic DNA extracts for the remaining taxa were diluted to 2.5 ng/ µl in 20 µl water. The Nextera Illumina library preparation kit (Illumina, San Diego, CA, USA) was used to prepare libraries for sequencing and the DNA Clean and Concentrator kit (Zymo Research, Irvine, CA, USA) was used for library sample purification. Sequencing was performed with the HiSeq 2000 instrument at the Iowa State University DNA Sequencing Facility, Ames, USA using single reads. This method produced 100 bp fragments. See Table 7 for details on sequencing techniques for each respective taxon.

All reads were first quality filtered using DynamicTrim v2.1 from the SolexaQA software package (Cox et al., 2010) with default settings, and then sequences less than 25 bp in length (default setting) were removed with LengthSort v2.1 in the same package.

Plastome assembly, annotation, and alignment

Plastome assembly was performed with entirely de novo methods. The Velvet software package (Zerbino and Birney, 2008) was run iteratively by loading previously assembled contigs into the Velvet assembler multiple times (see Wysocki et al., 2014 for details) with kmer lengths ranging from 19–85 bp increasing by steps of 6 bp. Contigs were scaffolded using the anchored conserved region extension (ACRE) method (Wysocki et al., 2014). Because Velvet and other de Bruijn graph-based programs cannot build across repeated areas (such as the inverted repeat regions in angiosperm plastomes), the plastome is assembled in

segments, which need to be manually joined. The number of contigs scaffolded for each taxon is reported in Table 7. Any remaining gaps in the plastomes were resolved using contigs or reads by locating overlapping regions of at least 20 bp that had zero mismatches and started at one end of the read or contigs. Paired-end reads that were used to resolve gaps were verified by checking the position and orientation of their downstream mate. Gaps were resolved until the circular map was complete with no gaps or ambiguities. Overlapping regions were identified and gap closure was performed using Geneious Pro (Biomatters Ltd., Auckland, NZ). A final assembly assessment was performed by mapping each set of reads to their respective plastome and locating any sequence inconsistencies as described in Wysocki et al. (Wysocki et al., 2014). Assembled plastomes were annotated by aligning to a closely-related and previously annotated reference plastome in Geneious Pro and transferring the annotations from the reference to the assembled plastome when the annotation shared a minimum similarity of 70%. The banked plastomes from *Arundinaria gigantea* (NC020341), *Bambusa oldhamii* (NC012927) and *Cryptochloa strictiflora* (JX235348) were used as annotation references for members of the tribes Arundinarieae, Bambuseae and Olyreae respectively.

A PCR experiment was performed to verify putative mitochondrial insertions in the *Pariana radiciflora* and *Eremitis* sp. plastomes. Two pairs of primers were used to amplify fragments in which mitochondrial sequence was found adjacent to plastid sequence. Two primers were designed based on the mitochondrial insert sequence in *Eremitis* sp. A BLASTn search of these two primer sequences showed 96-100% nucleotide identity to the *Ferrocalamus rimosivaginus* mitochondrial genome and no significant similarity to banked

chloroplast sequences. The other pair of primers were chosen based on flanking chloroplast sequence. Each pair of primers included one that annealed inside the insertion and one that annealed outside. The amplified fragments spanned insertion termini. The three designed primers were: 5′-GGGTCTCATCTGAAGGGAGGCAGGC-3′, 5′-GTGAGGCAGGTTCTCATGGTTCGG-3′ and 5′-GTGCTATCGGATCGGGTGAATTAGAG-3′, and the IRb 3 F primer from Dhingra and Folta (Dhingra and Folta, 2005) was also used. Amplifications were performed using the Fidelitaq system (Affymetrix, Santa Clara, CA) following the manufacturer's protocol. Products were separated electrophoretically on an agarose gel system.

Plastomes were arranged, beginning at the 5′ end, with the large single copy region (LSC) followed by the inverted repeat region B (IRb), ending with the short single copy region (SSC). Inverted repeat region A was omitted from the matrix to be used for phylogenomic analysis to prevent overrepresentation of the inverted repeat sequence. The assembled plastomes were then aligned, along with 14 previously published bambusoid plastomes and one pooid and one ehrhartoid grass plastome each, using the MAFFT alignment software (Katoh et al., 2005). The alignment was then inspected for structural mutations and adjusted manually to preserve tandem repeat boundaries and identify inversions. Regions that contained inversion mutations were deleted to remove false homology inferences. To test for potential differences in phylogenetic signal, all protein coding sequences were extracted from the alignment and concatenated for partitioned analyses.

Phylogeny estimation

Nucleotide positions that contained one or more gaps introduced by the alignments were omitted from the matrix. The Akaike Information Criterion (AIC) was used in the jModelTest software package v 2.1.3 (Guindon and Gascuel, 2003; Darriba et al., 2012) to compare models of character evolution. The General Time Reversible model of substitution, incorporating invariant sites and a gamma distribution (GTR + I + G), was among a group of equally best fit models (found in the 100% confidence interval) and was used in subsequent plastome analyses. Maximum likelihood analysis was performed using the RAxML v 8.0.5 software package (Stamatakis, 2006) with 1,000 non-parametric bootstrap replicates. Outgroup choice for Bambusoideae is complicated by the fact that divergence and radiation of the BEP subfamilies, possibly combined with undocumented extinctions, puts any candidate outgroup for Bambusoideae on comparatively long branches in phylogenetic trees (Wu and Ge, 2012) . This creates the potential for introducing phylogenetic artifacts. Full plastomes from the ehrhartoid grass *Zizania aquatica* (this paper) and the pooid grass *Lolium perenne* (NC009950) were included in the matrix as outgroup taxa. Non-parametric bootstrap values were generated using the Consense function of the Phylip software package (Felsenstein, 2005). An alternate topology was tested for the complete plastome partition in the likelihood framework. A second ML analysis was performed constraining the woody species to be monophyletic specifying identical parameters in the RAxML software. Constrained and unconstrained analyses were compared using the Shimodaira-Hasegawa (SH) test function included in PAUP* (Swofford, 2003). MrBayes 3.2.2 (Ronquist and Huelsenbeck, 2003) was used to perform a Bayesian inference analysis. The Markov chain

Monte Carlo (MCMC) analysis was run for 2 X 10,000,000 generations. Average standard deviation of split frequencies remained below 0.001 after the fifty percent burn-in. A neighbor-net analysis was then performed on the full plastome alignment to visualize character state conflict using the SplitsTree4 v. 4.13.1 (Huson and Bryant, 2006).

RESULTS

Assembly and alignment of plastomes

Read and contig assembly yielded complete plastomes for 18 bamboos and one ehrhartoid grass. Plastome lengths ranged from 135,320—143,810 base pairs (bp). Lengths of each plastome region are reported in Table 6. A multi-plastome sequence alignment was 132,707 bp in length after excluding one of the major inverted repeat (IR) regions. Removal of alignment columns containing gaps reduced the alignment length to 97,593 bp. The sequence alignment containing only protein coding regions was 54,548 bp in length and 52,941 bp after removal of gapped positions. See Table 7 for more information on sequencing techniques and results.

Unique plastome features

Plastomes are highly conserved chromosomes in which gene content, structure, and arrangement are quite similar across Poaceae (Bortiri et al., 2008). When infrequent events such as large insertion/deletion (indel) mutations or inversions do occur, they take on greater significance because of their rarity and therefore higher chance of indicating shared ancestry. Four of these were observed here among bambusoid plastomes, in each case marking a single synapomorphic event in our phylogeny:

1) A 2,706 bp insertion exclusive to sampled members of the Parianinae (*Eremitis* sp. and *Pariana radiciflora*) was found in the *rpl23-ndhB* intergenic spacer of the *Pariana radiciflora* plastome, while the *Eremitis* sp. plastome possessed this insertion plus an additional 1,242 bp inserted on the 3′ end, giving the insertion a total length of 4,938 bp (Fig. 3). A query of the NCBI nucleotide database using BLASTn (Altschul et al., 1997) revealed the highest scoring hit to be mitochondrial sequence from *Ferrocalamus rimosivaginus*, a member of the Arundinarieae (98-99% nucleotide identity). Subsequent BLAST hits were all of monocot mitochondrial origin. To confirm that this putative mitochondrial insertion is not the effect of an assembly artifact a PCR experiment was designed to amplify a region of approximately 2,400 bp by priming within and upstream of the insertions in both Parianinae. A second pair of primers was designed to amplify a region of similar size by priming within and downstream of the insertion in *Eremitis* sp. Note that this downstream region was not present in *P. radiciflora*. Amplification of these regions in *Eremitis* sp. produced two products that were both approximatelty 2,400 as expected. Amplification of the upstream region of *P. radiciflora* also showed a 2,400 bp product while the downstream region yielded no amplification, again as expected. 2) A deletion of 1,500 bp unique to the represented members of the subtribe Guaduinae (*Guadua weberbaueri, Olmeca reflexa, Otatea acuminata*) is also located in the same intergenic spacer (Fig. 3). 3) The alignment also revealed a 150 bp inversion in the *trnD-psbM* intergenic spacer exclusive to all sampled members of the subtribe Olyrinae (*Cryptochloa strictiflora, Diandrolyra* sp., *Lithachne pauciflora, Olyra latifolia, Raddia brasiliensis*). 4) An insertion in the *rps16-trnQ* intergenic spacer of approximately 500 bp was located in all members of Arundinarieae sampled in this

study (*Acidosasa purpurea*, *Arundinaria appalachiana*, *A. gigantea*, *A. tecta*, *Ferrocalamus rimosivaginus*, *Indocalamus longiauritus*, *Phyllostachys edulis*, *P. nigra*, *P. propinqua*, *Thamnocalamus spathiflorus*).

Full plastome phylogenomic analysis

Phylogeny estimation of full plastome sequences using maximum-likelihood (ML) and Bayesian inference (BI) generated trees with identical topologies. An annotated phylogenomic tree that includes all of the taxa can be found in Figure 4. All nodes were supported in the BI analysis with a posterior probability of 1.0. These trees supported monophyly of Arundinarieae, Bambuseae, and Olyreae with Bambuseae forming a well-supported sister relationship with Olyreae. Note that this is unlikely to be an artifact of long-branch attraction because the long-branch Olyreae associated with short-branch Bambuseae rather than the long-branch outgroup taxa. The Shimodaira-Hasegawa (SH) test (Shimodaira and Hasegawa, 1999; Goldman et al., 2000) allowed us to reject the alternative hypothesis of a monophyletic Bambuseae + Arundinarieae for the trees produced from these complete plastome sequences ($p < 0.001$). The Bambuseae diverged into two well-supported monophyletic lineages that represent neotropical and paleotropical woody bamboos. The neotropical bamboos segregated into two well-supported lineages, Chusqueinae (*Chusquea spectabilis*, *C. liebmannii*) and Guaduinae. The two representative species of Chusqueinae produced longer branch lengths than the rest of the woody bamboos with terminal branch lengths five times greater than those of the rest of Bambuseae.

The paleotropical bamboos displayed shorter branch lengths and lower support between the two sampled subtribes, Hickeliinae (*Hickelia madagascariensis*) and Bambusinae (82% ML bootstrap support). Bambusinae formed two well-supported clades: (*Dendrocalamus latiflorus* + *Bambusa* spp., and *Neololeba atra* + *Greslania* sp.). The genus *Bambusa* was resolved as monophyletic with 83% ML bootstrap support with very short branches and one internal node with 81% ML bootstrap support (Fig. 4). The Olyreae lineage demonstrated maximal support for all nodes with *Buergersiochloa bambusoides* sister to Olyrinae + Parianinae, both of which are monophyletic groups. Olyreae also displayed substantially longer branches than Bambuseae with mean internal branch lengths 5.5 times greater and mean terminal branch lengths 3.8 times greater (Fig. 4). Arundinarieae was resolved into two well-supported clades: ([*Arundinaria* spp. + *Acidosasa purpurea*] + *Ferrocalamus rimosivaginus* and [*Phyllostachys* spp. + *Indocalamus longiauritus*] + *Thamnocalamus spathiflorus*). *Arundinaria* was strongly supported as monophyletic (95% ML bootstrap support) with maximum support for intrageneric relationships. *Phyllostachys* was maximally supported as monophyletic yet exhibited less intrageneric support (78% ML bootstrap value) among the three species (Fig. 4).

Phylogenetic analysis of protein-coding regions

Maximum-likelihood and Bayesian analyses of protein-coding regions showed nearly identical topologies to the full plastome analysis, including a strongly supported Bambuseae + Olyreae. However, two differences at shallow nodes in the topology are present. The protein-coding analysis places *Dendrocalamus latiflorus* in a position embedded within the

genus *Bambusa* contrasting with the sister relationship of *D. latiflorus* and *Bambusa* recovered from the full plastome analysis. *Ferrocalamus rimosivaginus* exhibits a sister relationship to the rest of the Arundinarieae, which differs from its placement sister to the *Acidosasa + Arundinaria* clade in the full plastome analysis. Additionally, seven previously recovered nodes are supported at lower ML bootstrap and posterior probabilities.

DISCUSSION

Plastome tree topology

The monophyletic tribes, subtribes and genera retrieved here largely confirm those identified in previous studies. Notably, the plastome tree topology demonstrates paraphyly of the entire woody bamboo syndrome and suggests two independent origins of these characters or a common origin of the syndrome followed by its loss in Olyreae. The topology within Olyreae is well-supported, which can be attributed, in part, to its higher substitution rate and increased informative sites that are likely caused by the short generation times of this annually flowering lineage (Gaut et al., 1997). The New World Chusqueinae also exhibited a higher substitution rate in our ML analyses. While some species within Chusqueinae flower as infrequently as once every 70 years, flowering intervals are extremely variable in this lineage (Janzen, 1976; Clark, 1989; Guerreiro and de Agrasar, 2013). *Chusquea spectabilis* was formerly classified within the genus *Neurolepis*, which has shorter flowering intervals correlating with higher altitude habitats (Judziewicz et al., 1999). Although the phenology of *Chusquea liebmannii* is not well known, its higher substitution rates suggest that it may also flower relatively frequently. The substitution rates of the annually flowering outgroups

Lolium perenne (Skøt et al. 2011) and *Zizania aquatica* (http://eol.org/pages/1114723/overview) are also elevated and support the relationship between frequent reproduction and high substitution rates in the BEP clade (Gaut et al., 1997).

The topology of Olyreae in our tree agrees well with current taxonomy (Soreng et al. 2015). The three recognized subtribes (Olyrinae, Parianinae, and Buergersiochloinae) are sampled here and resolved as monophyletic groups with maximum support in our phylogenomic analyses. The deep divergence of *Buergersiochloa bambusoides* is of note. Olyreae have a contemporary distribution in the New World except *Olyra latifolia* which, though largely Neotropical is also widespread in Africa/Madagascar. Another exception is *B. bambusoides*, which is endemic to New Guinea. The biogeography of Olyreae argues for a New World origin and radiation followed by limited long-distance dispersals. However, the position of *B. bambusoides* as sister to the remaining Olyreae recovered in our analysis and many others contradicts this hypothesis. Our topology suggests an Old World origin followed by a New World dispersal and radiation with a long distance dispersal event for *O. latifolia* to Africa/Madagascar, likely via birds feeding on the pseudo-berries produced by this species (Judziewicz et al., 1999). Other historical scenarios are more complicated invoking repeated dispersals and extinction events and are difficult to reconcile with the phylogenomic topology presented here.

Low bootstrap support and short branch lengths that obscure intrageneric relationships within *Bambusa* can be attributed to the relatively close evolutionary

relationships of these species, which is reflected in high sequence similarity accompanied with a weak phylogenetic signal. Plastomes from this genus share high sequence similarity (99.8%) and fewer intrageneric synapomorphic mutations. The possibility of intrageneric hybridization as well as hybridization events between closely related genera soon after their divergence also presents an issue when the exact branching order of these species is considered (Wong and Low, 2011; Goh et al, 2013). The long generation time of the Bambuseae could allow artifacts of hybridization to persist well after their divergences.

Greslania includes three or four species endemic to New Caledonia with similarities of reproductive morphology to *Hickelia* (Dransfield, 2002). The genus is taxonomically associated with the broadly-defined Bambusinae (Soreng et al. 2015), but its phylogenetic position is somewhat more specifically defined by Chokthaweepanich (2014) as sister to what is called the CDMNPPS (*Cyrtochloa-Dinochloa-Mullerochloa-Neololeba-Parabambusa-Pinga-Sphaerobambos*) clade. Our well-supported phylogenomic placement of *Greslania*, sister to *Neololeba atra* (Fig. 4; 100% ML bootstrap support), is consistent with the previous work. The Australasian distribution of these taxa offers additional insights. The tectonic history of New Caledonia placed it in longstanding isolation from Australia for some 50 to 65 Ma (Coleman, 1980). Some endemic plants of New Caledonia are late Cretaceous relicts, suggesting a distribution resulting from long-standing historical vicariance (Lowry, 1998). However, evidence of more recent geologic history of total submergences supports a contrasting view, that the New Caledonian flora can be no older than 37 Ma (Grandcolas et al., 2008). The phylogenomic position of *Greslania*, embedded within the relatively young clade of paleotropical woody bamboos, which have an estimated age

ranging from 19.6 to 25 million-years (Bouchenak-Khelladi et al, 2010; Ruiz-Sanchez, 2011), is consistent with the recent geological evidence and suggests a mid-Tertiary long-distance dispersal from a *Neololeba*-like ancestral taxon. Further plastome sequencing among the CDMNPPS clade will be required to further refine the position of *Greslania*.

Horizontal gene transfer between organellar genomes and other unique plastome features

The full plastome sequence assemblies revealed a 2,706 bp insertion of mitochondrial DNA in the *rpl23-ndhB* intergenic spacer within the IR region of *Pariana radiciflora* and a nearly identical insertion (99.6% identity) in the same region from the closely-related *Eremitis* sp., with an additional 2,232 base pairs appended to the 3′ end giving the insert a total length of 4,938 bp. A sequence similarity search using BLAST matched a 3,191 bp fragment of the insertion in *Eremitis* sp. to a region of the *rps7-atp6* intergenic spacer in the mitochondrial genome of *Ferrocalamus rimosivaginus*, a member of the Arundinarieae. Nearly the entire insert in *P. radiciflora* showed significant sequence similarity to the same region within the *F. rimosivaginus* mitochondrial genome. Although a mitochondrial fragment from Olyreae should exhibit high similarity to a mitochondrial genome from within the same lineage, only two mitochondrial genomes have been sequenced among the Bambusoideae to date (*Bambusa oldhamii* and *F. rimosivaginus*). Because of the rarity of this type of intergenomic transfer (Rice and Palmer, 2006; Goremykin et al., 2008; Straub et al., 2013), several tests were conducted to verify accurate plastome assembly in this region. Note that this putative insertion was originally assembled de novo using Velvet (a de Bruijn graph assembler) in both taxa. The insertion was embedded in contigs of 25.9 and 42.8 kbp

in *Eremitis* sp. and *P. radiciflora* respectively. Mapping the two sets of reads to their respective assemblies that include the insertion produced a continuum of perfectly overlapping reads that spanned the entire hypothesized mitochondrial insertions. The upstream and downstream boundaries of each insertion exhibited coverage of 14- and 16-fold respectively in *Eremitis* sp., and 124 and 107 respectively in *P. radiciflora*, which compare favorably with the overall coverage of each plastome (15.4 and 119.1, respectively). Mapping each set of reads to their respective flanking sequences produced regions identical to those of the insertion with no sign of consistent mismatching or misassembly, nor were there any clear patterns of coverage interruption when approaching each locus. Although it could be expected that this area would show higher coverage due to the reference mapping using reads from both the plastid and mitochondrial regions, the inserts show comparable coverage to the rest of the genome. The mitochondrial inserts are represented sufficiently to produce an assembly but are not proportionally represented in the read pool. One possible explanation could be that the base composition in the mitochondrial inserts is less AT rich than in the rest of the plastome (55% compared to 61%) as the PCR step during Illumina library preparation has been shown to introduce base composition bias in fragments (Aird et al., 2011). Another possibility is that this region was deleted from the mitogenome in the course of transfer to the plastome.It should also be noted that the Illumina libraries for *P. radiciflora* and *Eremitis* sp. were prepared using different methods (TruSeq Nano and Nextera respectively), sequenced at different facilities using paired-end and single-end, respectively, and plastomes were assembled independently using completely *de novo*

methods. Finally, the insert was confirmed with a PCR experiment using plastome/mitogenome primer pairs.

We hypothesize that this event originated from a recombination between the plastome and the homologous regions within the mitochondrial genome most probably in a common ancestor of these two taxa. The appearance of this mitochondrial insertion in two species of Parianinae is striking, and most parsimoniously interpreted as a single event even though one of the inserts is 2.2 kbp longer than the other. Given the rarity of mitochondrial insertions in grass plastomes, two such similar events in closely related taxa is more difficult to explain than a single insertion with subsequent differential degradation of this noncoding DNA. Events in which a mitochondrial genome incorporates DNA sequences of plastome origin are not rare, especially in bamboos (Ma et al., 2012). This creates homologous regions between the cytoplasmic organelles, which following further mitochondrial rearrangements might facilitate recombination of additional mitochondrial sequences into the chloroplast. While the mitochondrial genomes have yet to be sequenced in *P. radiciflora* and *Eremitis* sp., querying the mitochondrial genome of *Bambusa oldhamii* with its own plastome sequence using BLASTn reveals over forty regions of significant sequence similarity longer than 100 bp in length. The much less frequent horizontal gene transfer from mitochondrion to plastid has been observed and verified in other plant species (Rice and Palmer, 2006; Goremykin et al., 2008; Straub et al., 2013).

The subset of Olyreae that possess the unique 150 bp inversion in the *trnD-psbM* intergenic spacer includes representatives of only five genera of Olyrinae. The Olyrinae are

well-supported as monophyletic in this study as is also suggested in Oliveira et al. (2014), which indicates that this inversion likely occurred once in the common ancestor of this lineage. The high number of substitutions and indels accumulated between species within this unique inversion either supports the notion that the inversion event occurred early in the history of this lineage or that mutation rates are elevated. An imperfect eight bp inverted repeat flanking the inverted region (CCYTTTTY -inversion- GAAAAAGG) suggests that a possible inversion mechanism could be a stem-loop formation induced recombination.

Conclusions

This study successfully characterizes the full plastome sequences of 16 tropical bamboos, one temperate bamboo, and one ehrhartoid grass. Three sequences from Guaduinae and two from Chusqueinae represent the first completely assembled plastome sequences from the New World tropical woody lineage. Though full plastome sequences have been assembled from Old World tropical species (Zhang et al., 2011; Wu et al., 2009) our taxonomic sampling of this lineage extends beyond the *Bambusa* - *Dendrocalamus* clade. This study also marks the first full plastome phylogenomic analysis to be performed within Olyreae. Two plastomes from Olyreae reveal the first evidence of a synapomorphic mitochondrial-to-plastid horizontal gene transfer in monocots.

This phylogenomic study supports paraphyly of the woody bamboo syndrome. However, the scope of the relationships presented here is restricted to the maternally inherited evolutionary signal, which demonstrates considerable conflict in the phylogenomic network analysis at the node where the three main lineages diverge (Fig. 5). A study on three single-copy nuclear

markers performed by Triplett et al. (2014) outlined a scenario in which the extant allopolyploid woody bamboos are a result of two separate hybridization events between at least four distinct precursor lineages. Herbaceous bamboos were supported in a sister relationship to a progenitor lineage that eventually diversified into precursor lineages that hybridized to form the extant woody bamboos. The Triplett et al. (Triplett et al., 2014) study clarified some of the complexities of bamboo diversification and provided evidence that the apparent paraphyly of the woody syndrome in bamboos may be an artifact of analysis with exclusively plastid loci. However, note that one out of the three nuclear markers potentially supported the robust tropical woody-herbaceous bamboo sister relationship in plastid studies by embedding the diploid herbaceous clade within lineages exclusive to tropical woody bamboos. Further study using a wider variety of nuclear markers may clarify this significant event in bamboo diversification.

A comparative study on the transcriptomics of the lignin biosynthesis and deposition pathways could provide further insight on the evolution of the woody character. A single origin of characters found in the woody bamboo syndrome would be supported by similar expression profiles between Bambuseae and Arundinarieae in the genes for enzymes and transcription factors involved in lignin biosynthesis and deposition and formation of bisexual florets while differing expression profiles could suggest otherwise. Other potential expansions on this study are an examination of the phylogenetic signals given by other molecular characters such as mitochondrial coding sequences and microstructural changes.

Table 6. NCBI nucleotide database accession numbers and lengths of regions and subregions for plastomes analyzed in this study

Taxon	Tribe	Total length	LSC [a]	SSC [b]	IR [c]	Accession	Voucher
Acidosasa purpurea	Arundinarieae	139,697	83,273	12,834	21,795	NC015820	N/A
Arundinaria appalachiana	Arundinarieae	139,547	83,222	12,717	21,804	NC023934	N/A
Arundinaria gigantea	Arundinarieae	138,935	82,632	12,709	21,797	NC020341	N/A
Arundinaria tecta	Arundinarieae	139,499	83,161	12,730	21,804	NC023935	N/A
Ferrocalamus rimosivaginus	Arundinarieae	139,467	83,091	12,718	21,829	NC015831	N/A
Indocalamus longiauritus	Arundinarieae	139,668	83,273	12,811	21,792	NC015803	N/A
Phyllostachys edulis	Arundinarieae	139,679	83,213	12,870	21,798	NC015817	N/A
Phyllostachys nigra	Arundinarieae	139,839	83,234	12,879	21,863	NC015826	N/A
Phyllostachys propinqua	Arundinarieae	139,704	83,228	12,878	21,799	NC016699	N/A
Thamnocalamus spathiflorus	Arundinarieae	139,498	83,310	12,594	21,797	KJ871005	LC 1319 (ISC)

Taxon	Tribe	Total length	LSC [a]	SSC [b]	IR [c]	Accession	Voucher
Bambusa arnhemica	Bambuseae	139,287	82,790	12,901	21,798	KJ870989	PMP 1846 (CAN)
Bambusa bambos	Bambuseae	142,772	79,972	12,868	24,966	KJ870988	BI 1
Bambusa emeiensis	Bambuseae	139,491	82,976	12,911	21,802	NC015830	N/A
Bambusa oldhamii	Bambuseae	139,347	82,889	12,878	21,790	NC012927	N/A
Chusquea liebmannii	Bambuseae	138,001	81,501	12,892	21,804	KJ871001	LC & LA 1710 (ISC)
Chusquea spectabilis	Bambuseae	136,848	80,743	12,671	21,717	KJ870990	XL & LC 919 (ISC)
Dendrocalamus latiflorus	Bambuseae	139,369	82,975	12,884	21,755	NC013088	N/A
Greslania sp.	Bambuseae	139,264	82,581	12,979	21,852	KJ870993	GM (MO)
Guadua weberbaueri	Bambuseae	135,320	82,803	12,929	19,794	KP793062	XL & MK 582 (TULV)
Hickelia madagascariensis	Bambuseae	138,276	81,925	12,743	21,804	KJ870994	SD 1349 (K)
Neololeba atra	Bambuseae	139,395	82,905	12,926	21,782	KJ870996	LC & JT 1663 (ISC)
Olmeca reflexa	Bambuseae	136,213	82,726	12,945	20,271	KJ870997	Francisco Botanical Garden 312 (GCR)
Otatea acuminata	Bambuseae	136,351	82,859	12,948	20,272	KJ871003	LC & WZ 1348 (ISC)

Taxon	Tribe	Total length	LSC [a]	SSC [b]	IR [c]	Accession	Voucher
Buergersiochloa bambusoides	Olyreae	138,122	81,746	12,856	21,760	KJ871000	SD 1365 (Kew)
Cryptochloa strictiflora	Olyreae	134,332	80,554	12,766	20,506	JX235348	N/A
Diandrolyra sp.	Olyreae	137,469	81,752	13,259	21,229	KJ870991	LC 1301 (ISC)
Eremitis sp.	Olyreae	143,810	80,984	13,232	24,797	KJ870992	LC & WZ 1343 (ISC)
Lithachne pauciflora	Olyreae	135,385	79,465	13,676	21,122	KJ871002	LC 1297 (ISC)
Olyra latifolia	Olyreae	135,834	80,642	12,770	21,211	KF515509	N/A
Pariana radiciflora	Olyreae	139,650	81,847	13,221	22,291	KJ871004	LC & WZ 1344 (ISC)
Raddia brasiliensis	Olyreae	135,739	80,713	13,000	21,013	KJ870998	LC & LA 1713 (ISC)
Zizania aquatica	Oryzeae (Ehrhartoideae)	136,354	82,009	12,587	20,879	KJ870999	JS 20870 (CAN)
Lolium perenne	Poeae (Pooideae)	135,246	80,000	12,428	21,409	NC009950	N/A

Table 7. Sequencing details for all plastomes newly assembled for this study.

Taxon	Tribe	Number of reads	Library preparation method	Sequencing method	Mean coverage	Number of scaffolded contigs
Thamnocalamus spathiflorus	Arundinarieae	7,098,663	TruSeq	Single-end	54.8	5
Bambusa arnhemica	Bambuseae	2,292,120	TruSeq	Single-end	25.7	15
Bambusa bambos	Bambuseae	5,279,202	TruSeq	Single-end	53.2	3
Chusquea liebmannii	Bambuseae	23,707,569	TruSeq Nano	Paired-end	126.8	6
Chusquea spectabilis	Bambuseae	7,348,756	Nextera	Single-end	23.7	9
Greslania sp.	Bambuseae	13,881,568	Nextera	Single-end	142.1	3
Guadua weberbaueri	Bambuseae	29,431,971	Nextera	Single-end	94.9	9
Hickelia madagascariensis	Bambuseae	13,509,970	Nextera	Single-end	43.7	10
Neololeba atra	Bambuseae	28,569,106	TruSeq Nano	Paired-end	497.3	13
Olmeca reflexa	Bambuseae	5,400,472	Nextera	Single-end	51.3	6
Otatea acuminata	Bambuseae	14,532,488	TruSeq Nano	Paired-end	134.9	4
Buergersiochloa bambusoides	Olyreae	12,592,122	Nextera	Single-end	124.8	6

Taxon	Tribe	Number of reads	Library preparation method	Sequencing method	Mean coverage	Number of scaffolded contigs
Diandrolyra sp.	Olyreae	10,004,619	Nextera	Single-end	100	4
Eremitis sp.	Olyreae	4,674,178	Nextera	Single-end	15.4	13
Lithachne pauciflora	Olyreae	14,773,417	Nextera	Single-end	233.1	4
Pariana radiciflora	Olyreae	23,398,974	TruSeq Nano	Paired-end	119.6	4
Raddia brasiliensis	Olyreae	6,828,240	Nextera	Single-end	40.1	3
Zizania aquatica	Oryzeae (Ehrhartoideae)	6,018,945	TruSeq	Single-end	66.3	3

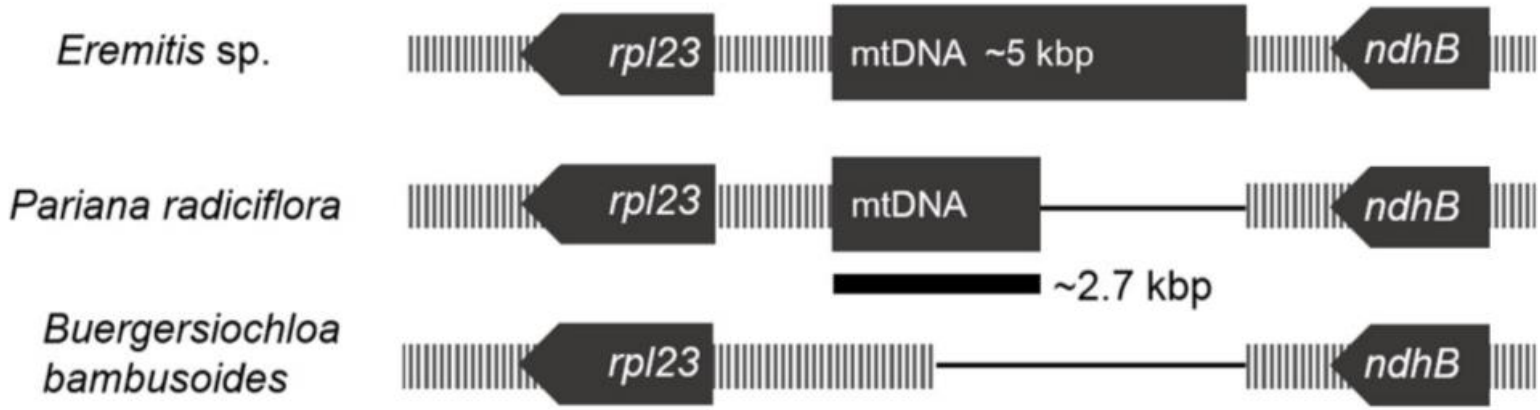


Figure 3. Relative positions of putative mitochondrial insertions in the *Pariana radiciflora* and *Eremitis* sp. plastomes. A diagram of the region in *Buergersiochloa bambusoides* is also included to illustrate an example of a typical grass plastome without the insertion. Solid bars represent relative gene positions, striped bars represent intergenic regions and thin lines represent gaps that were introduced to preserve downstream alignment. Note that this figure is not drawn to scale.

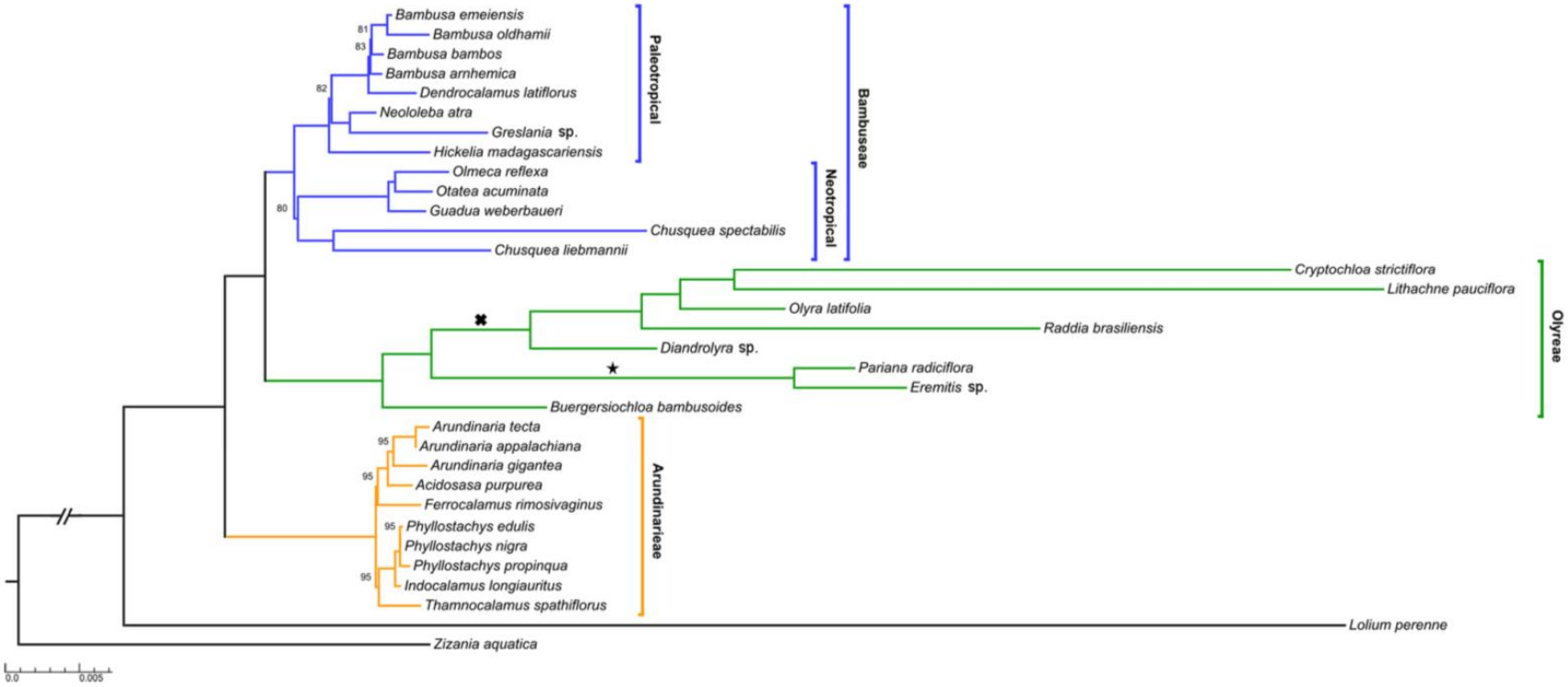


Figure 4. Maximum likelihood phylogram for all complete plastomes. Branch lengths are given in substitutions per site. The star indicates the hypothesized origin of the mitochondrion-to-plastid horizontal gene transfer event. The cross indicates the hypothesized origin of the 150 bp inversion in subtribe Olyrinae. Nodes are supported at a 100% maximum likelihood bootstrap score unless reported. All nodes were supported with a posterior probability of 1.0.

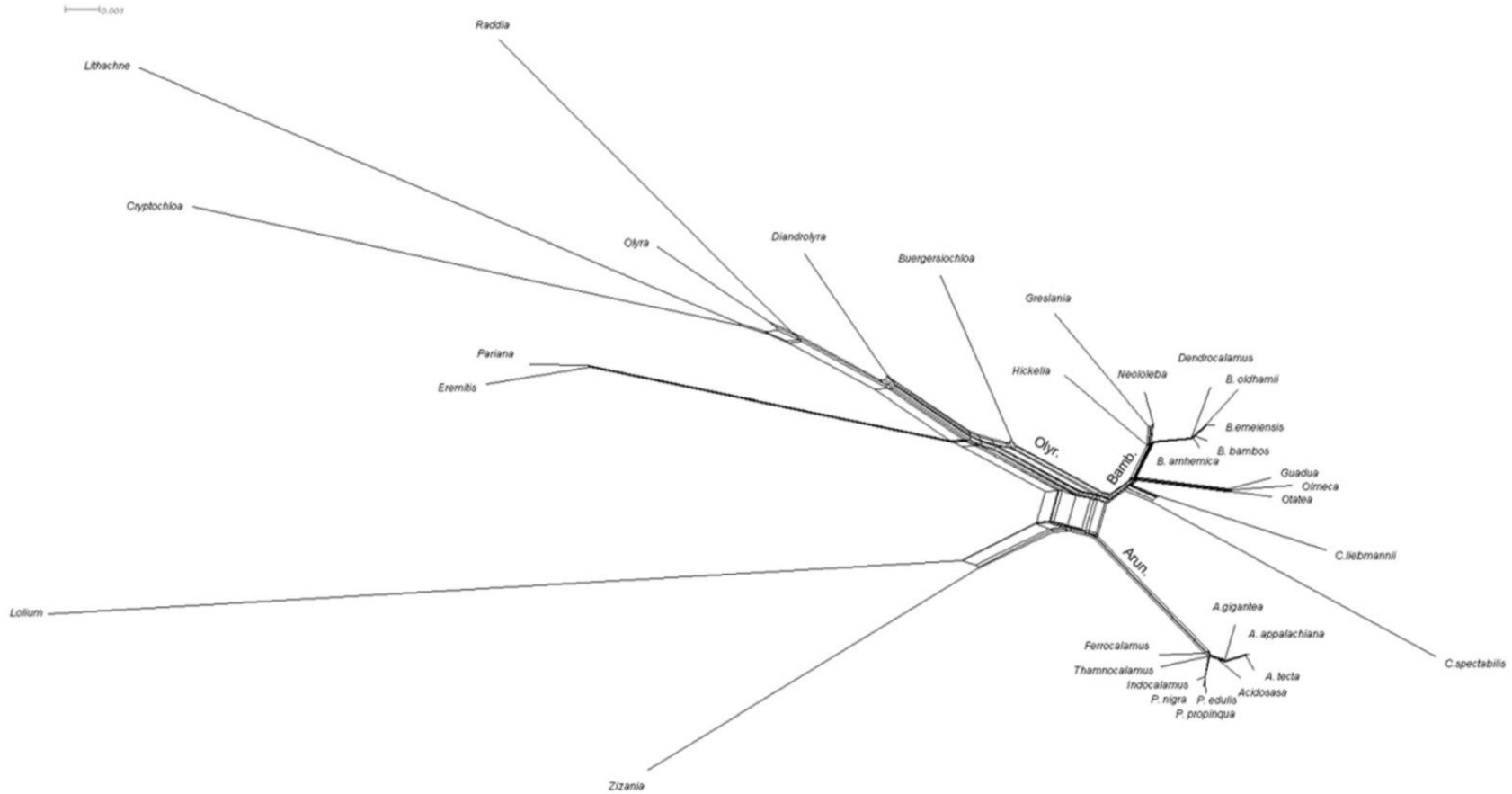


Figure 5. A neighbor-net analysis indicates conflicting phylogenetic signals in the data. The three main bamboo lineages are indicated. Note the branches for outgroup taxa *Lolium* and *Zizania* were truncated to facilitate visibility. Bamb.: Bambuseae; Olyr.: Olyreae; Arun.: Arundinarieae

CHAPTER 3

PHYLOGENY OF THE BAMBUSOIDEAE: REVISITED

ABSTRACT

The previous chapter outlines the basics of full plastome bambusoid phylogeny. However there is much to be learned from an increase in sampling. Here, 53 bamboo plastomes were sequenced in full using next-generation sequencing. Of these new plastomes, 34 represent those from newly sequenced species while 19 come from species that had been represented in full plastome studies before. Maximum likelihood and Bayesian analysis largely supported previous taxonomic designations in the lineages of Arundinarieae and the subtribes of Olyreae. The subtribes of Bambuseae exhibited some discordance with previous taxonomic designations. Additionally, a neighbor-net analysis was performed to visualize character conflict and propose alternative hypotheses. With relationships at a variety of levels left unresolved or poorly supported, further sampling in nuclear markers will be necessary to clarify the evolutionary history of the bamboos.

INTRODUCTION

Full plastid genomes (plastomes) have been used to clarify phylogenetic relationships among the Bambusoideae for the past seven years. Plastomes from *Bambusa oldhamii* and *Dendrocalamus latiflorus* were the first to be fully sequenced (Wu et al., 2009). Two years later, the first bambusoid plastomes were sequenced using next-generation

sequencing from *Bambusa emeiensis* and five temperate woody species (Zhang et al., 2011), which allowed them to increase support for interspecific relationships. Wu and Ge (2012) sequenced one additional plastome from *Phyllostachys propinqua*, as well as plastomes from other grass subfamilies, and used full plastomes to support the hypothesis in which Bambusoideae is sister to the subfamily Pooideae. Burke et al. (2012) sequenced the first plastome from an herbaceous bamboo (*Cryptochloa strictiflora*) and used divergence estimations to make phylogeographic inferences on the dispersal of *Arundinaria gigantea*. Burke et al. (2014) then went on to sequence two additional plastomes from *Arundinaria* and an additional herbaceous species (*Olyra latifolia*).

The plastome phylogenomics of the temperate woody clade was explored with further sampling by Ma et al. (2014) and Attigala et al. (in review), sequencing eighteen and ten new plastomes, respectively, from Arundinarieae. Further work by Wysocki et al. (2015) balanced sampling among the three tribes adding plastomes from 13 more tropical bamboos, including both woody and herbaceous species. These plastome data showed maximum support for the monophyly of the tropical clade (Bambuseae + Olyreae). Also in this study the first example of a mitogenome to plastome horizontal transfer of DNA in monocots was discovered between species in two herbaceous sister genera. Ma et al. (2015) subsequently were able to repeat the results of Wysocki et al. (2015), dutifully replicating the sequencing from these taxa and also extending a previously documented Guaduinae-specific deletion in the distantly related *Otatea glauca*.

The goal of this study is to expand current understanding of bambusoid phylogeny using a wider representation of full plastomes. New species from the Bambuseae were

represented in this study including 11 species from Bambusinae, three species from Melocanninae, three species from Guaduinae and two species from Chusqueinae. Although a subtribal classification has not been designated for Arundinarieae, 11 new plastomes from temperate woody bamboos were generated. The fewest new plastomes were generated from Olyreae in which one new herbaceous species was represented. A total of 19 full plastomes from 17 species with previously sequenced plastomes were assembled in duplicate for this study to verify previously published relationships.

METHODS

Full plastome assembly

Illumina reads for 53 samples, each corresponding to a different species of bamboo, were provided by Dr. Domitille Chalopin, Department of Genetics, University of Georgia, Athens. These reads were originally generated for a project on transposable elements within the Bambusoideae. Reads were sequenced paired-end and downloaded from BaseSpace (http://basespace.illumina.com). The plastomes sequenced in this study fall into two categories: 35 were from species that were to be newly sequenced for this study and 18 were from species that were previously sequenced and deposited into Genbank (Table 8). Read files were filtered for plastid homology by extracting reads that mapped to previously published bamboo plastomes.

Reads were then assembled de novo using SPAdes v. 3.5.0 (Bankevich et al., 2012) with kmer lengths of 19--85 increasing in steps of six. Assembled contigs were then scaffolded using the anchored conserved regions extension method (Wysocki et al.,

2014). This method used conserved plastid sites of over 20 bp in length to order contigs, which were extended using reads and contigs until overlap of at least 20 bp was encountered in an adjacent contig. When the full circular chromosome maps were completed, plastome subregions were identified using the methods of Burke et al. (2012). Each plastome was then verified by mapping the raw reads onto the original plastome and manually repairing inconsistencies between the reads and the assembled plastome. Plastomes were annotated by aligning each to a previously published reference plastome from a closely related species and transferring annotations that had sequence identity of over 70%. This could be performed because grass plastomes are completely collinear. Protein coding sequences were adjusted to preserve reading frames, start codons and stop codons.

Alignment and phylogenomic analyses

The 53 newly sequenced plastomes were combined with 66 published bambusoid plastomes and two outgroup plastomes downloaded from GenBank. Plastomes were aligned using MAFFT (Katoh and Standley, 2013). Nucleotide positions in which one or more gaps were introduced by the alignment were eliminated from the matrix to minimize the inclusion of ambiguously aligned sites. One of the inverted repeat regions was also eliminated from the matrix to prevent duplication of phylogenetic signal.

Maximum-likelihood (ML) phylogeny estimation was performed using RAxML v. 8.2.4 (Stamatakis, 2014). A GTR plus I plus Γ model of rate heterogeneity was implemented, which was used as an appropriate model in this subfamily by Wysocki et al. (2015). A bootstrap analysis with 1,000 nonparametric bootstrap pseudoreplicates was

also performed. A Bayesian inference (BI) analysis was performed using MrBayes v. 3 (Ronquist and Huelsenbeck, 2003). Dirichlet priors were used for base frequencies and the rate matrix. Uniform priors were used for the shape parameter (ALPHA), proportion of invariable sites (I) and topology. A Markov chain Monte Carlo (MCMC) analysis was run to completion for 2 X 10,000,000 generations with four chains each. The pooid grass *Lolium multiflorum* (NC_019651) and oryzoid grass *Zizania aquatica* (NC_026967) were included as outgroup taxa in both of the phylogenetic analyses. To visualize character conflict, a neighbor-net analysis was performed using SplitsTree4 (Huson and Bryant, 2006) on the alignment that had been stripped of gapped positions and had one inverted repeat region removed.

RESULTS

Full plastome assembly

A total of 53 plastomes were assembled completely for this study. Names, tribal designations and collection sites are reported in Table 8. After removal of one inverted repeat region, the full plastome alignment was 140,571 in length and reduced to 95,761 bp after removal of gapped positions.

Phylogeny estimation

Using both maximum-likelihood and Bayesian methods, Bambusoideae and all three main tribes were fully supported as monophyletic with full support for the Arundinarieae

forming a sister relationship to a Bambuseae + Olyreae clade. Annotated ML trees that detail each tribe are shown in Fig. 6.

The three subtribes of Olyreae were recovered as monophyletic with Olyrinae + Parianinae forming a sister relationship to the monotypic Buergersiochloinae. The newly sequenced *Raddia distichophylla* formed a monophyletic relationship with the two other accessions of *Raddia* (two representatives of *R. brasiliensis*). All genera and duplicate species were recovered as monophyletic. All branches in this clade were supported at 100% maximum likelihood bootstrap value (MLBV) and posterior probability (PP) of 1.00.

Representatives of all described tropical woody subtribes, except for Racemobambosinae, were included in this study. Neotropical subtribes (Arthrostylidiinae, Guaduinae, and Chusqueinae) and paleotropical subtribes (Bambusinae, Hickeliinae, and Melocanninae) formed clades with 51% MLBV; 0.62 PP and 100% MLBV; 1.00 PP support respectively. The three neotropical woody subtribes formed clades that were all maximally supported with Arthrostylidinae+Guaduinae in a sister relationship to Chusqueinae. The newly sequenced *Guadua angustifolia*, *G. paniculata* and *Otatea aztecorum* were recovered as monophyletic with congeneric members. All species of *Chusquea* were recovered as a monophyletic group. One of the newly sequenced representatives of *C. circinata* formed a sister relationship to the conspecific plastome from GenBank. However, the other representative from *C. circinata* formed a sister relationship to *C. liebmannii*, which was supported at 100%

MLBV but only 0.71 PP. *Chusquea gigantea* and the two representatives of *C. culeou* formed a clade but the two conspecific taxa were fully supported as paraphyletic.

Each of the paleotropical subtribes was not recovered completely as monophyletic. Melocanninae were recovered as monophyletic with *Cephalostachyum pergracile* sister to *Schizostachyum brachycladum* + *C. scandens* (99% MLBV; 1.00 PP). The two representatives of *Neohouzeaua mekongensis* were not included in this clade. Hickeliinae (*Hickelia madagascariensis* + *Nastus elatus*) was also recovered and maximally supported as monophyletic. Bambusinae were recovered as monophyletic except for *Neololeba atra* and *Greslania* sp., which were sister to the Hickeliinae (100% MLBV; 1.00 PP). The remaining Bambusinae were in two well-supported clades; one of which included all representative members of *Bambusa*. Conspecific representatives of three *Bambusa* species were each individually recovered as monophyletic although the genus was not. *Gigantochloa hasskarliana* was sister to the *B. multiplex* clade and *Oxytenathera abyssinica* was sister to the *B. atra* + *B. arnhemica* clade.

The temperate woody clade has not yet been classified into subtribal groups. However, Attigala et al. (2014; in review) and Zeng et al. (2010) classified Arundinarieae into 12 distinct lineages. Lineage XI (*Ampelocalamus calcareus*) was sister to the remaining taxa. Subtending lineage XI were two large clades. The first clade comprised lineages IV, VI, and VIII; all of which were maximally supported, but inter-lineage relationships were unresolved.

The second clade comprised lineages I, II, III, V, VII, IX. The deepest diverging subclade included lineages I + VII, which were paraphyletic. Two representatives of

Bergbambos tessellata (I) were paraphyletic with *Thamnocalamus spathiflorus* (VII), but this relationship had low support (53% MLBV; 0.60 PP). Another subclade comprised lineages II, III, IX and XII. All were supported maximally as monophyletic but inter-lineage relationships were unresolved. Lineage V fell sister to this subclade. Intra-lineage relationships fell into a large range of support levels. The subclade that included both representatives of *Ampelocalamus scandens,* which was recovered as paraphyletic with *Drepanostachyum khasianum* either with low support (66% MLBV) or was not resolved (BI analysis). A well supported subclade, sister to the aforementioned subclade, included all members of *Phyllostachys,* most members of *Fargesia* (except *F. yunnanensis*), *Arundinaria fargesii* and *Pseudosasa cantorii*. *Phyllostachys* was recovered as monophyletic with strong support (99% MLBV; 1.00 PP) and intrageneric relationships were not well supported or resolved.

DISCUSSION

Plastome phylogeny estimation

A tropical bamboo clade (Olyreae+Bambuseae) was recovered in this study as sister to the temperate woody bamboo clade (Arundinarieae) with strong support. This relationship has been published previously in studies that used plastid markers for phylogenetic analyses (Kelchner et al., 2013; Wysocki et al., 2015). In addition to a parallel gain of the woody character in Arundinarieae and Bambuseae or the gain and loss of the woody character in Olyreae, other traits, such as bisexual floral structure and flowering phenology, would have had to evolve along with the parallel evolution of

woodiness. Recent studies that have used nuclear markers for phylogenetics (Triplett et al., 2014; Wysocki et al., in review) have recovered a woody clade. Triplett et al. (2014) also hypothesized a single origin of woody bamboos that resulted from a hybridization of diploid progenitor genomes that likely shared phenotypic traits with Olyreae.

The relationships recovered within the herbaceous bamboo clade (Olyreae) were consistent with previous studies (Oliveira et al., 2014; Wysocki et al., 2015) and robust. This could be attributed to greater phylogenetic information, reflected in their branch lengths, which were generally greater than branch lengths within both woody bamboo clades. Branch lengths correlate with a higher rate of substitutions (Gaut et al., 1997) and may signify a lower chance for hybridization events. Note that the neighbor-net analysis (Fig. 7c) produced an almost tree-like pattern for these taxa. Moreover, this study added only one new species to the full plastome representation of the herbaceous bamboos.

The neotropical woody bamboos (Bambuseae), which comprise Chusqueinae, Arthrostylidiinae and Guadinae, were very weakly supported as monophyletic in this study. The plastome sampling in this study actually decreased support from 80% MLBV and 1.00 PP (Wysocki et al., 2015) to 57% MLBV and 0.62 PP, which implied that additional molecular information actually obscured the relationships. The alternative hypothesis is that Chusqueinae forms a sister relationship to the paleotropical woody bamboos + Arthrostylidiinae + Guaduinae (Kelchner et al., 2013). Both of these possibilities are observable in the neighbor-net analysis (Fig. 7b) in which the Guaduinae+Arthrostylidiinae clade is in a midpoint position between the Chuqueinae and

paleotropical woody bamboo clade. Besides the deep node, all previously described subtribes and genera were robustly supported as monophyletic and fully resolved.

The paleotropical woody bamboos (Bambuseae) were maximally supported as monophyletic, but contained several internal divergences with weak support and conflict with previous taxonomic designations. Two genera within the monophyletic Melocanninae (*Cephalostachyum* + *Schizostachyum*) are paraphyletic. *Neohouzeaua mekongensis* (two accessions), which is classified in Melocanninae, were paraphyletic with Bambusinae. This placement is incongruent with the robust placement of this genus in the Melocanninae in a six plastid locus phylogeny (Chokthaweepanich, 2014).

Bambusinae and Hickeliinae composed the crown group of the paleotropical bamboos. Hickeliinae were placed within the Bambusinae in a sister relationship to *Neololeba* + *Greslania*. Although this placement is not congruent with single evolutionary origins of subtribes, it was supported by Kelchner et al. (2013) in a similar phylogenetic position. Sister to this group are two clades of Bambusinae. Wysocki et al. (2015) sampled full plastomes in these clades so that one was represented with *Bambusa oldhamii*, *B. emeiensis*, *B. arnhemica*, and *B. bambos* and the other sampled only with *Dendrocalamus latiflorus*. This produced the appearance of recovering monophyletic sister genera; however, increasing the sampling in both clades demonstrated that this was not the case. In this study the *Bambusa* clade also contained one species each from *Gigantochloa* and *Oxytenanthera*. The *D. latiflorus* clade also contained one species from *Gigantochloa*, two species from *Dinochloa*, the two previously mentioned

representatives of *N. mekongensis,* and two additional species from *Dendrocalamus*. Both clades were well-supported.

Arundinarieae was represented by 11 out of 12 previously described lineages. Missing from this study are representatives from the *Indocalamus sinicus* (X) lineage. Sequence information from this species that was previously published in Genbank (KJ531442) is a partial plastome, which is missing over 55% of the full plastome. All previously described lineages have been recovered as monophyletic except for the *B. tessellata* (I) and *T. spathiflorus* (VII) lineages. The two representatives from *B. tessellata* form a very poorly supported paraphyletic relationship with *T. spathiflorus* (52% MLBV; 0.06 PP). This likely results from the three plastomes being very similar to each other. The lineages that were recovered as monophyletic were maximally supported but relationships between them were poorly supported or unresolved. The neighbor-net analysis illustrates this with each lineage exhibiting a distinct origin with a highly reticulated network where the origins meet (Fig 7a).

Duplicated plastome sequences

All species that were sequenced in duplicate, except four, exhibited relationships sister to the respective previously published plastome. The newly sequenced *Bergbambos tessellata* branched in a sister relationship to the previously published *Thamnocalamus spathiflorus + Bergbambos tessellata* clade. However, all three plastomes exhibited over 99.9% sequence similarity and the node uniting the two previously published plastomes was weakly supported. Although these two species possess two distinct morphologies and

were classified into two major clades (I and VII), recent hybridizations, in nature and in cultivation, could be responsible for their nearly identical plastome sequences. Misidentification is also a possibility because identification of bamboo species using only vegetative characteristics can prove difficult. The two representative members of *Ampelocalamus scandens* were in a paraphyletic relationship with *Drepanostachyum khasianum* with low MLBV support. Since the three plastomes are in the same lineage (V) and exhibit 99.99% sequence identity, this could be also be a hybridization issue.

The two newly sequenced and one previously published representative of *Chusquea circinata* also formed a paraphyletic relationship with its congener *C. liebmannii*. These sequences exhibited 99.8% sequence identity. The two representatives from *C. culeou* formed a well-supported paraphyletic relationship with *C. gigantea*. These three newly sequenced plastomes also exhibited 99.8% sequence identity. Although these relationships could be explained with hybridizations, they could also have resulted from the relatedness of these congeneric species and the relatively short amount of time they have had to accumulate mutations that would differentiate plastome sequences.

Conclusions

The difficulties in recovering monophyletic genera and subtribes within both woody bamboos clades are likely multifaceted. Plastid-based phylogenies recover only the signal passed maternally, which does not allow hybridizations to be detected. The long generation times between episodic flowering in bamboos allow for fewer chances of

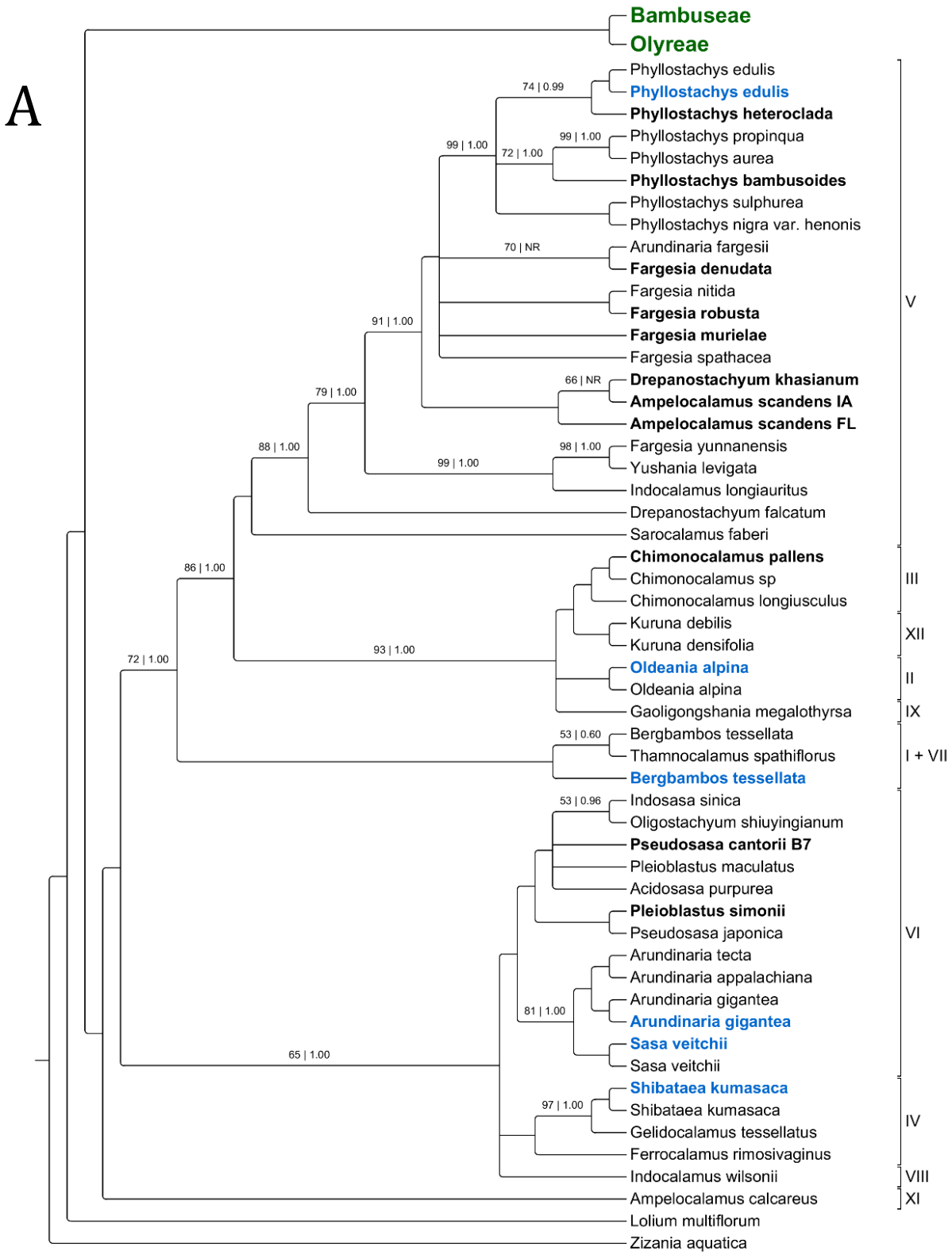
nucleotide substitutions between recently diverged taxa and a lower likelihood for a reproductive barrier to form. This could result in hybridizations and phylogenetic conflict between the maternally-inherited plastid genes and nuclear genes. Nuclear genes tend to be responsible for morphological characteristics, many of which are responsible for taxonomic designations. This naturally puts taxonomic classification systems at odds with the phylogenetic signal generated from the plastid. Phylogenetic analyses are difficult to produce confidently using nuclear sequences due to the presence of multiple gene copies and the complications that arise with orthology assessment. While nuclear phylogenetic analyses may not recover monophyletic genera in complicated taxonomic systems, they do provide additional insights on the origin of these complications.

Table 8. Collection sites for samples used to generate full plastome sequences in this study.

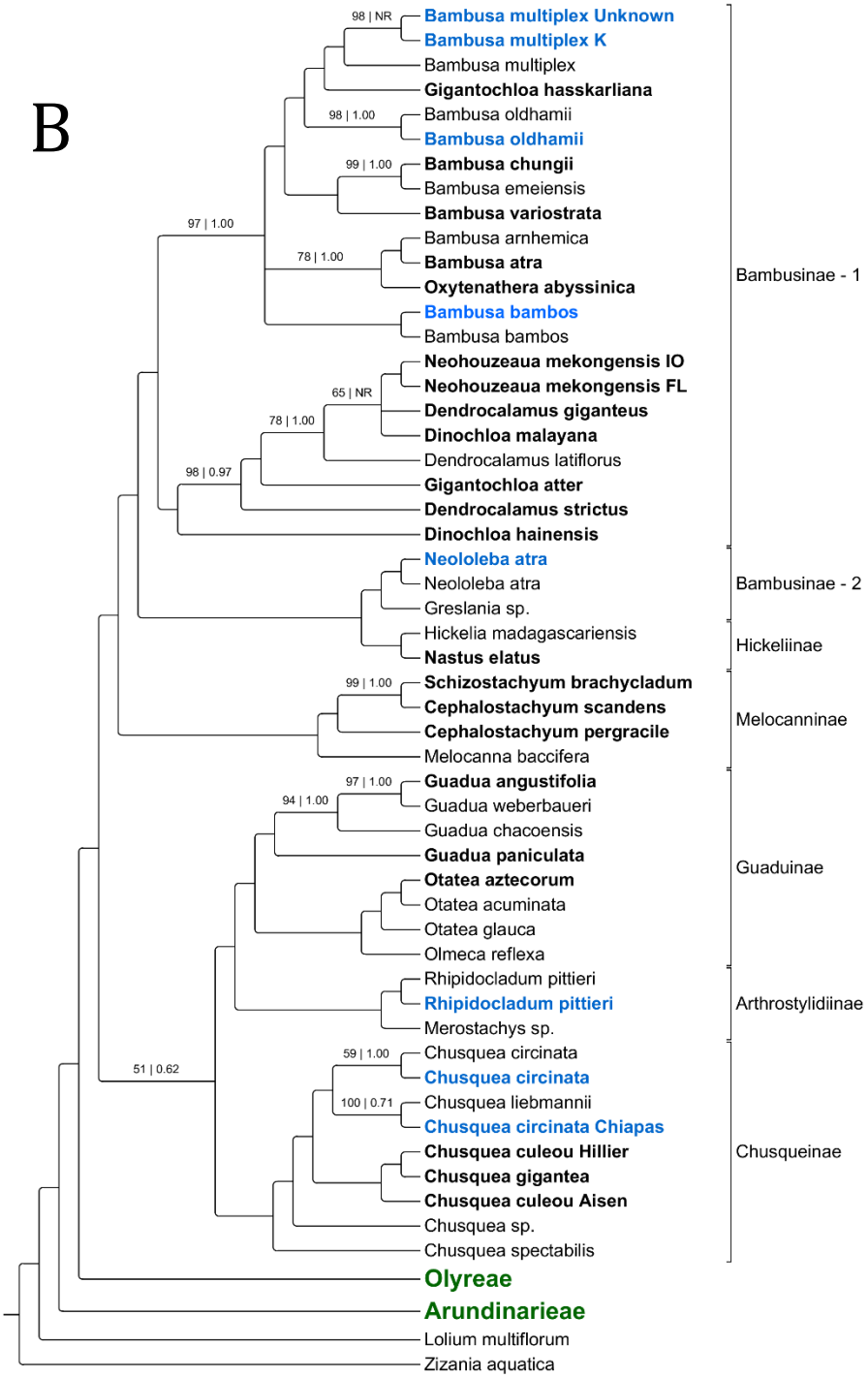
Taxon	**Tribe**	**Collection Site**
Ampelocalamus scandens IO	Arundinarieae	Iowa State University, Ames, IA
Ampelocalamus scandens FL	Arundinarieae	Fairchild Tropical Garden, Coral Gables, FL
Chimonocalamus delicatus	Arundinarieae	Fairchild Tropical Garden, Coral Gables, FL
Drepanostachyum khasianum	Arundinarieae	Fairchild Tropical Garden, Coral Gables, FL
Arundinaria gigantea	Arundinarieae	Bamboo Garden Nursery, North Plains, OR
Bergbambos tessellata	Arundinarieae	Bamboo Garden Nursery, North Plains, OR
Chimonocalamus pallens	Arundinarieae	Bamboo Garden Nursery, North Plains, OR
Fargesia murielae	Arundinarieae	Bamboo Garden Nursery, North Plains, OR
Fargesia denudata	Arundinarieae	Bamboo Garden Nursery, North Plains, OR
Fargesia robusta campbell	Arundinarieae	Bamboo Garden Nursery, North Plains, OR
Phyllostachys bambusoides	Arundinarieae	Bamboo Garden Nursery, North Plains, OR
Phyllostachys heteroclada	Arundinarieae	Bamboo Garden Nursery, North Plains, OR
Phyllostachys edulis Jaquith	Arundinarieae	Bamboo Garden Nursery, North Plains, OR
Pleioblastus simonii	Arundinarieae	Bamboo Garden Nursery, North Plains, OR
Sasa veitchii	Arundinarieae	Bamboo Garden Nursery, North Plains, OR
Shibataea kumasaca	Arundinarieae	Bamboo Garden Nursery, North Plains, OR
Oldeania alpina	Arundinarieae	Tradewinds Bamboo Nursery, Gold Beach, OR
Pseudosasa cantorii B-7	Arundinarieae	USDA-ARS Fruit and Nut Tree Station, Byron, GA
Otatea aztecorum	Bambuseae	Iowa State University, Ames, IA
Rhipidocladum pittieri	Bambuseae	Iowa State University, Ames, IA
Neohouzeaua mekongensis IO	Bambuseae	Iowa State University, Ames, IA
Oxytenanthera abyssinica	Bambuseae	Iowa State University, Ames, IA
Guadua angustifolia	Bambuseae	Fairchild Tropical Garden, Coral Gables, FL

Bambusa atra	Bambuseae	Fairchild Tropical Garden, Coral Gables, FL
Bambusa bambos	Bambuseae	Fairchild Tropical Garden, Coral Gables, FL
Bambusa chungii	Bambuseae	Fairchild Tropical Garden, Coral Gables, FL
Bambusa multiplex K	Bambuseae	Fairchild Tropical Garden, Coral Gables, FL
Bambusa multiplex Unknown	Bambuseae	Fairchild Tropical Garden, Coral Gables, FL
Bambusa oldhamii	Bambuseae	Fairchild Tropical Garden, Coral Gables, FL
Cephalostachyum pergracile	Bambuseae	Fairchild Tropical Garden, Coral Gables, FL
Cephalostachyum scandens	Bambuseae	Fairchild Tropical Garden, Coral Gables, FL
Dendrocalamus giganteus	Bambuseae	Fairchild Tropical Garden, Coral Gables, FL
Dendrocalamus strictus	Bambuseae	Fairchild Tropical Garden, Coral Gables, FL
Dinochloa malayana	Bambuseae	Fairchild Tropical Garden, Coral Gables, FL
Dinochloa hainensis	Bambuseae	Fairchild Tropical Garden, Coral Gables, FL
Gigantochloa atter	Bambuseae	Fairchild Tropical Garden, Coral Gables, FL
Gigantochloa hasskarliana	Bambuseae	Fairchild Tropical Garden, Coral Gables, FL
Nastus elatus	Bambuseae	Fairchild Tropical Garden, Coral Gables, FL
Neohouzeaua mekongensis FL	Bambuseae	Fairchild Tropical Garden, Coral Gables, FL
Neololeba atra	Bambuseae	Fairchild Tropical Garden, Coral Gables, FL
Schizostachyum brachycladum	Bambuseae	Fairchild Tropical Garden, Coral Gables, FL
Bambusa variostrata	Bambuseae	Fairchild Tropical Garden, Coral Gables, FL
Chusquea culeou Hillier	Bambuseae	Bamboo Garden Nursery, North Plains, OR
Chusquea culeou Aisen	Bambuseae	Bamboo Garden Nursery, North Plains, OR
Chusquea circinata	Bambuseae	Tradewinds Bamboo Nursery, Gold Beach, OR
Chusquea circinata Chiapas	Bambuseae	Tradewinds Bamboo Nursery, Gold Beach, OR
Chusquea gigantea	Bambuseae	Tradewinds Bamboo Nursery, Gold Beach, OR
Guadua paniculata	Bambuseae	Tradewinds Bamboo Nursery, Gold Beach, OR
Eremitis sp.	Olyreae	Iowa State University, Ames, IA
Pariana radiciflora	Olyreae	Iowa State University, Ames, IA

Diandrolyra sp.	Olyreae	Iowa State University, Ames, IA
Lithachne pauciflora	Olyreae	Iowa State University, Ames, IA
Raddia brasiliensis	Olyreae	Iowa State University, Ames, IA
Raddia distichophylla	Olyreae	Iowa State University, Ames, IA


A
Bambuseae
Olyreae
Phyllostachys edulis
Phyllostachys edulis
Phyllostachys heteroclada
Phyllostachys propinqua
Phyllostachys aurea
Phyllostachys bambusoides
Phyllostachys sulphurea
Phyllostachys nigra var. henonis
Arundinaria fargesii
Fargesia denudata
Fargesia nitida
Fargesia robusta
Fargesia murielae
Fargesia spathacea
Drepanostachyum khasianum
Ampelocalamus scandens IA
Ampelocalamus scandens FL
Fargesia yunnanensis
Yushania levigata
Indocalamus longiauritus
Drepanostachyum falcatum
Sarocalamus faberi
Chimonocalamus pallens
Chimonocalamus sp
Chimonocalamus longiusculus
Kuruna debilis
Kuruna densifolia
Oldeania alpina
Oldeania alpina
Gaoligongshania megalothyrsa
Bergbambos tessellata
Thamnocalamus spathiflorus
Bergbambos tessellata
Indosasa sinica
Oligostachyum shiuyingianum
Pseudosasa cantorii B7
Pleioblastus maculatus
Acidosasa purpurea
Pleioblastus simonii
Pseudosasa japonica
Arundinaria tecta
Arundinaria appalachiana
Arundinaria gigantea
Arundinaria gigantea
Sasa veitchii
Sasa veitchii
Shibataea kumasaca
Shibataea kumasaca
Gelidocalamus tessellatus
Ferrocalamus rimosivaginus
Indocalamus wilsonii
Ampelocalamus calcareus
Lolium multiflorum
Zizania aquatica
74 | 0.99
99 | 1.00
99 | 1.00
72 | 1.00
70 | NR
91 | 1.00
66 | NR
79 | 1.00
88 | 1.00
98 | 1.00
99 | 1.00
86 | 1.00
93 | 1.00
72 | 1.00
53 | 0.60
53 | 0.96
81 | 1.00
65 | 1.00
97 | 1.00
V
III
XII
II
IX
I + VII
VI
IV
VIII
XI

B


98 | NR
Bambusa multiplex Unknown
Bambusa multiplex K
Bambusa multiplex
Gigantochloa hasskarliana
98 | 1.00
Bambusa oldhamii
Bambusa oldhamii
99 | 1.00
Bambusa chungii
Bambusa emeiensis
Bambusa variostrata
97 | 1.00
Bambusa arnhemica
78 | 1.00
Bambusa atra
Oxytenathera abyssinica
Bambusa bambos
Bambusa bambos
Neohouzeaua mekongensis IO
Neohouzeaua mekongensis FL
65 | NR
Dendrocalamus giganteus
78 | 1.00
Dinochloa malayana
Dendrocalamus latiflorus
98 | 0.97
Gigantochloa atter
Dendrocalamus strictus
Dinochloa hainensis
Neololeba atra
Neololeba atra
Greslania sp.
Hickelia madagascariensis
Nastus elatus
99 | 1.00
Schizostachyum brachycladum
Cephalostachyum scandens
Cephalostachyum pergracile
Melocanna baccifera
97 | 1.00
Guadua angustifolia
94 | 1.00
Guadua weberbaueri
Guadua chacoensis
Guadua paniculata
Otatea aztecorum
Otatea acuminata
Otatea glauca
Olmeca reflexa
Rhipidocladum pittieri
Rhipidocladum pittieri
Merostachys sp.
59 | 1.00
Chusquea circinata
51 | 0.62
Chusquea circinata
100 | 0.71
Chusquea liebmannii
Chusquea circinata Chiapas
Chusquea culeou Hillier
Chusquea gigantea
Chusquea culeou Aisen
Chusquea sp.
Chusquea spectabilis
Olyreae
Arundinarieae
Lolium multiflorum
Zizania aquatica
Bambusinae - 1
Bambusinae - 2
Hickeliinae
Melocanninae
Guaduinae
Arthrostylidiinae
Chusqueinae

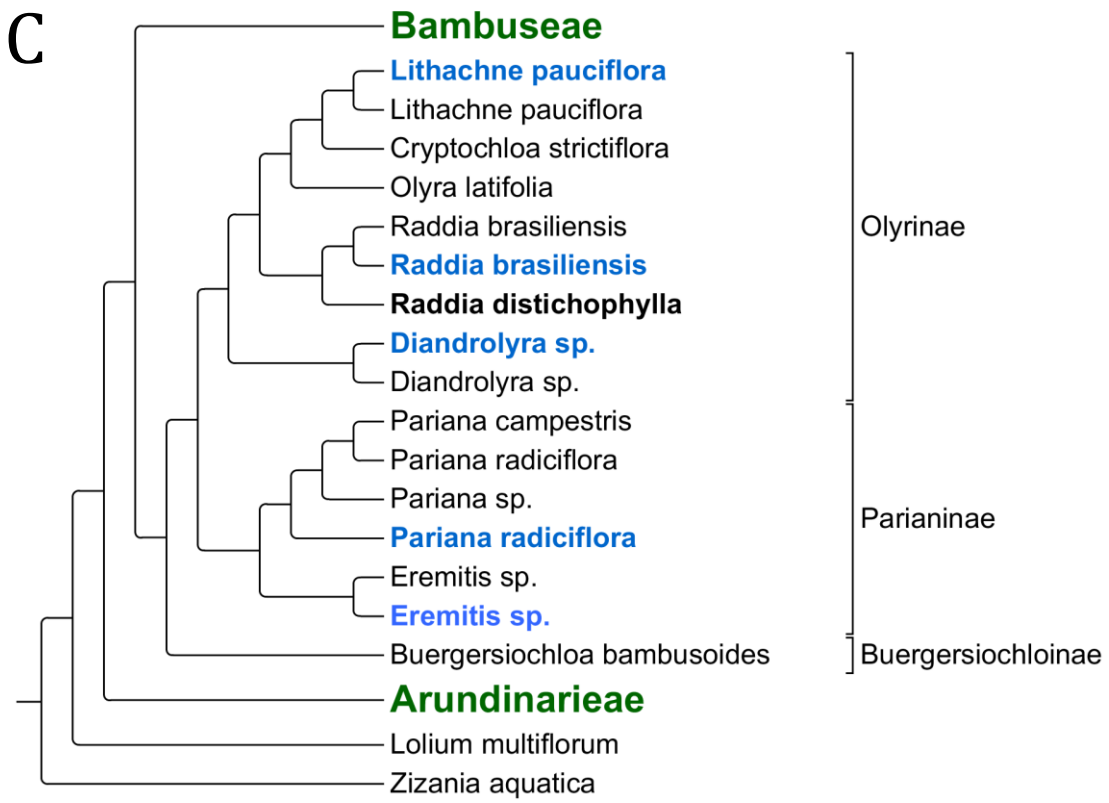
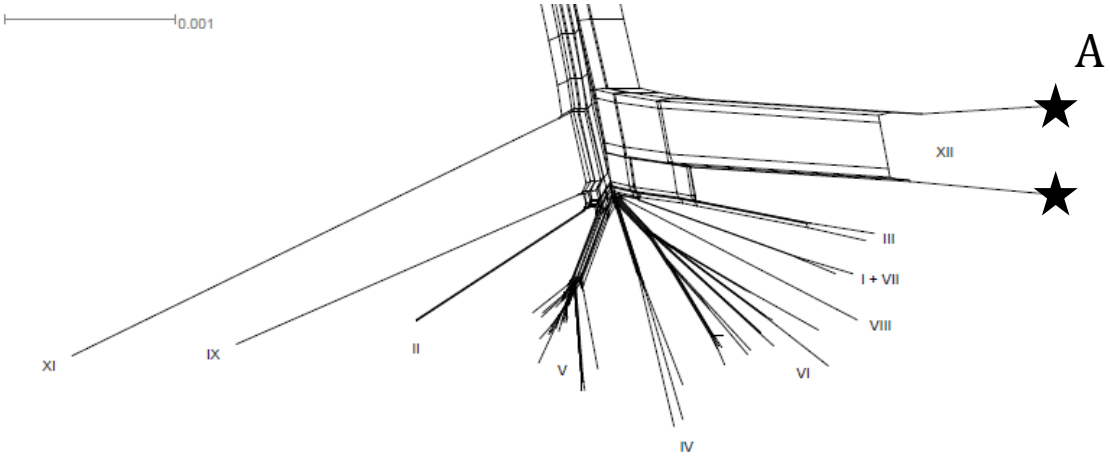
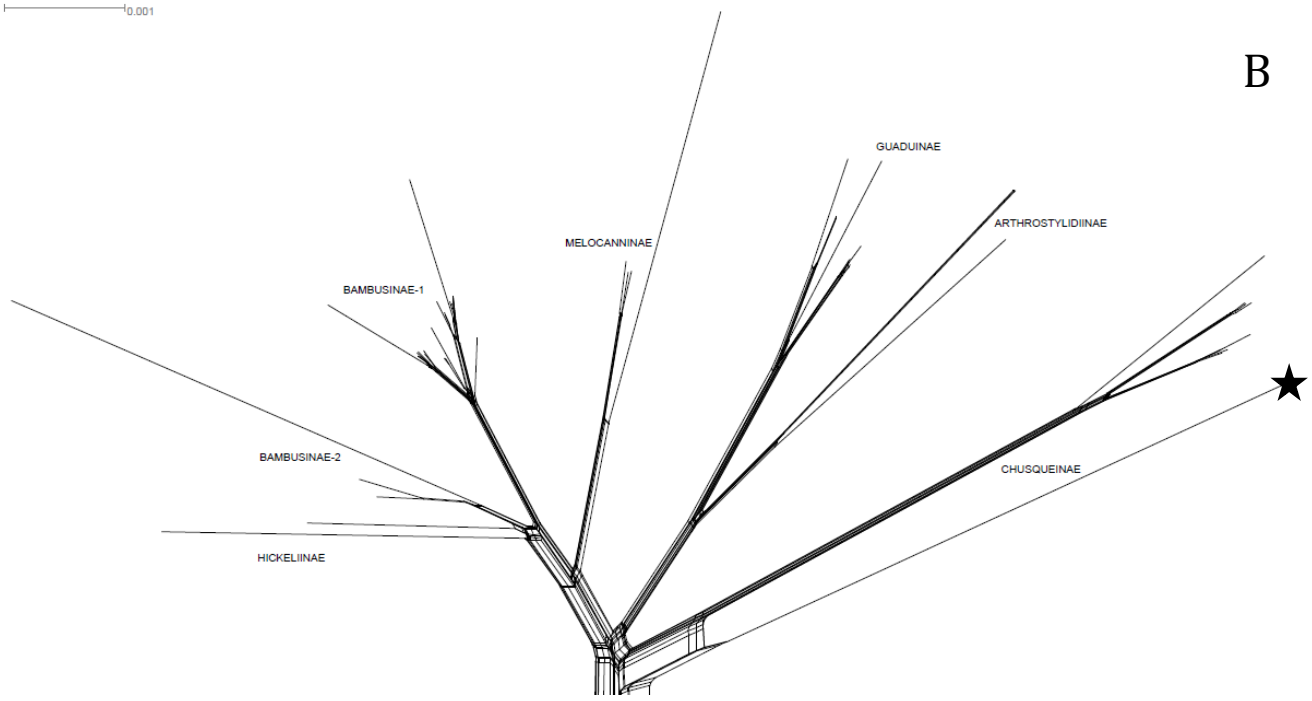


Figure 6. Three cladograms detailing relationships among Arundinarieae (A), Bambuseae (B) and Olyreae (C). Taxon names in **bold** font represent plastomes that were newly sequenced for this study. Plastomes that were previously sequenced in other studies and duplicated here are shown in blue. Support values represent maximum likelihood bootstrap support (MLBV) and posterior probabilities (PP). Unmarked nodes are supported at 100% MLBV and PP = 1.00. Branch lengths are redundant.


0.001
A
XII
III
I + VII
VIII
VI
IV
V
II
IX
XI


0.001
B
GUADUINAE
ARTHROSTYLIDIINAE
MELOCANNINAE
BAMBUSINAE-1
BAMBUSINAE-2
CHUSQUEINAE
HICKELIINAE

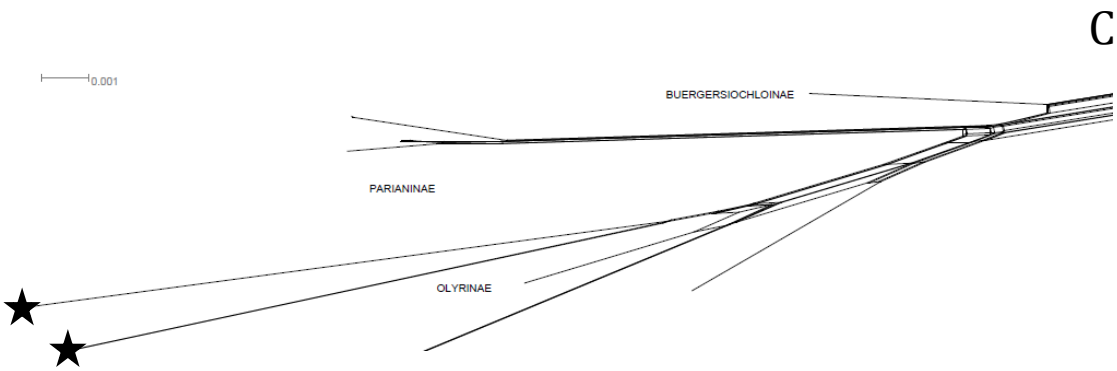


Figure 7. Neighbor-net analysis performed in SplitsTree4 for Arundinarieae (A), Bambuseae (B) and Olyreae (C). Subtribes and lineages are indicated. Stars indicate branches that continue longer but were truncated to conserve space.

CHAPTER 4

THE FLORAL TRANSCRIPTOMES OF FOUR BAMBOO SPECIES (BAMBUSOIDEAE; POACEAE): SUPPORT FOR COMMON ANCESTRY AMONG WOODY BAMBOOS.

ABSTRACT

Next-generation sequencing now allows for total RNA extracts to be sequenced in non-model organisms such as bamboos, an economically and ecologically important group of grasses. Bamboos are divided into three lineages, two of which are a woody perennials with bisexual flowers, which undergo gregarious monocarpy. Members of the third lineage, which are herbaceous perennials, possess unisexual flowers that undergo annual flowering events. Transcriptomes were assembled using both reference-based and de novo methods. These two methods were tested by characterizing transcriptome content using sequence alignment to previously characterized reference proteomes and by identifying Pfam domains. Because of the striking differences in floral morphology and phenology between the herbaceous and woody bamboo lineages, MADS-box genes, transcription factors that control floral development and timing, were characterized and analyzed in this study. Transcripts were identified using phylogenetic methods and categorized as A, B, C, D or E-class genes, which control floral development, or *SOC* or *SVP*-like genes, which control the timing of flowering events. Putative nuclear orthologues were also identified in bamboos to use as phylogenetic markers. Instances of

gene copies exhibiting topological patterns that correspond to shared phenotypes were observed in several gene families including floral development and timing genes. Alignments and phylogenetic trees were generated for 3,878 genes and for all genes in a concatenated analysis. Both the concatenated analysis and those of 2,412 separate gene trees supported monophyly among the woody bamboos, which is incongruent with previous phylogenetic studies using plastid markers.

INTRODUCTION

The Bambusoideae are a subfamily of perennial forest grasses endemic to every continent except Europe and Antarctica and comprise approximately 1,450 species (Bamboo Phylogeny Group (BPG), 2012; Kelchner et al., 2013). Bamboos exhibit a combination of characters that uniquely distinguish the subfamily within the larger evolutionary radiation of Poaceae. Bambusoideae is divided phylogenetically into three well-supported lineages: temperate woody (Arundinarieae), tropical woody (Bambuseae) and herbaceous (Olyreae) bamboos (Sungkaew et al., 2009; Kelchner et al., 2013). Both lineages of woody bamboos are characterized by complex rhizome systems, a tree-like habit with highly lignified and usually hollow culms, well-differentiated culm leaves, well developed aerial branching, foliage leaf blades with outer ligules, and bisexual spikelets. Woody bamboos typically exhibit gregarious flowering cycles followed by death of the parent plants (monocarpy) (Clark et al., 2015). They also serve as an economically important resource as they produce timber, fiber, food and other products. Herbaceous bamboos lack well differentiated culm leaves and outer ligules combined

with relatively weakly lignified culms, restricted vegetative branching, unisexual spikelets and seasonal flowering (Clark et al., 2015).

Despite the uncertainty of their phylogenetic relationships (Zhang 2000; Zhang and Clark, 2000; GPWG, 2001; Bouchenak-Khelladi et al., 2008; Sungkaew et al., 2009; Kelchner et al., 2013; Triplett et al., 2014), the two woody bamboo lineages share aspects of phenology and sexual systems suggestive of common ancestry. Their phenological patterns can be especially striking as they can exhibit extremely long intervals between flowering periods (3-120 years), which may be synchronized between disjunct populations (Janzen, 1976). The subsequent die-off following a flowering event can result in sudden ecological consequences such as lower shade levels in former bamboo forests (Marchesini et al., 2009).

Floral characteristics of Olyreae contrast with those of woody bamboos in both phenology and sexual systems, as the herbaceous species flower annually and possess unisexual spikelets, which are either segregated into different inflorescences or found together in a mixed inflorescence, in both cases on monoecious plants. Phenological differences between herbaceous and woody bamboos impact phylogenetic studies. Members of Olyreae generally exhibit elevated mutation rates compared to those of the woody bamboos, which are correlated with shorter generation times and longer branch lengths in phylogenetic trees (Gaut et al., 1997; Kelchner et al., 2013; Oliveira et al., 2014; Wysocki et al., 2015).

Floral development in angiosperms has been found to be largely controlled by MADS-box genes (Coen and Meyerowitz, 1991). Named after four of their homologues

(*MCM1*, *AGAMOUS*, *DEFICIENS*, *SRF*), these transcription factors control the development of each of the four floral whorls (sepal, petal, stamen, and carpel). The function of MADS box genes in floral development has been extensively studied using another grass, *Oryza sativa*, as well as the eudicots *Antirrhinum majus* and *Arabidopsis thaliana*. The mechanism of development can be generally described using the classical 'ABC' model (Coen and Meyerowitz, 1991). In this model, sepals develop with the expression of A-class genes, petals develop with the expression of A and B-class genes, stamens develop with the expression of B and C-class genes, and carpels develop with the expression of C-class genes. Subsequent studies have added D and E-class genes to the model in which expression of E-class genes is required for B and C-class function (Pelaz et al., 2000) and D-class gene expression is required for ovule development (Skinner et al., 2004; Dreni et al, 2007). A further refinement on this understanding of floral developmental genetics is the quartet model, which suggests that MADS-box proteins work in groups of four to initiate transcription (Theißen and Saedler, 2001).

In addition to development of floral structures, the expression of some MADS box genes can affect the timing of flowering. The *SUPRRESSION OF OVER-EXPRESSION OF CONSTANS 1* (*SOC1*) gene in *A. thaliana*, as well as its homolog *OsMADS56* in *O. sativa*, have been characterized as being involved in several steps in the process of inducing floral development (Ryu et al., 2009; Lee and Lee, 2010). The *SHORT VEGETATIVE PHASE* (*SVP*) gene in *A. thaliana* and *OsMADS22 +55* in *O. sativa* have been also shown to control flowering and can act as antagonists of the *SOC1*

genes (Hartmann et al., 2000; Lee et al., 2012). These genes are of particular interest in bamboos because of the aforementioned phenological characteristics.

Next-generation sequencing (NGS) has allowed gene expression to be examined at a large scale for non-model organisms using the RNA-Seq method (Wang et al., 2009). RNA-Seq typically uses mRNA selected by the poly-A tails to filter only for eukaryotic protein-coding transcripts. The RNA is reverse-transcribed into cDNA, which is fragmented and sequenced on a NGS platform such as Illumina. The RNA-Seq method was first used in a bamboo species by Zhang et al. (2012) to characterize the floral transcriptome of *Dendrocalamus latiflorus* (Bambuseae), with a follow up study on seed, leaf, stem, shoot and root tissue of the same species (Liu et al., 2012). Changes in transcript abundance in shoots of *Phyllostachys edulis* (Arundinarieae) during development were examined by (Peng et al., 2013a) and a similar study on flowers was conducted (Gao et al., 2014).

In this study, floral transcriptomes are characterized from four species representing three major bamboo lineages: *Guadua inermis* and *Otatea acuminata* (tropical woody Bambuseae), *Phyllostachys aurea* (temperate woody Arundinarieae) and *Lithachne pauciflora* (herbaceous Olyreae). The floral structures of these species vary as *L. pauciflora* has unisexual florets, while the three woody bamboos have bisexual florets. *P. aurea* and *O. acuminata* have three stamens (Ruiz-Sanchez et al., 2011; Kellogg, 2015b), while *G. inermis* has six. *G. inermis*, *O. acuminata*, and *L. pauciflora* have two stigmas (Kellogg, 2015a,c), while *P. aurea* has three (Kellogg, 2015b).

These transcriptomes were analyzed to meet three complementary objectives. First, the content of each transcriptome was characterized by transcript identification and categorization. Transcripts conserved across the grass family and other plants are identified as well as bamboo-specific transcripts using two methods of assembly. In the past two years, the first draft nuclear genome for a bamboo (*Phyllostachys heterocycla*) was sequenced and published (Peng, et al., 2013b). Our sampling scheme includes transcriptome data from the congeneric *P. aurea*, which allows for a reference-based assembly to be tested and compared to a de novo assembly.

Second, the evolutionary histories of the genes important to floral timing and development in grasses were explored. Because both bisexual and unisexual flowers and two very distinct phenological patterns are present within the taxa examined here, including the developmental A, B, C, D and E-class MADS-box genes and the *SVP* and *SOC1* gene families, which are involved in reproductive timing.

Lastly, the evolutionary history of the bamboo species examined here was examined by alignment and analysis of putative nuclear orthologues. Genes were selected based on their single-copy status in two reference taxa from the Bambusoideae-Oryzoideae-Pooideae (BOP) clade (*Brachypodium distachyon* and *O. sativa*). The goal of this portion of the study was to find markers from transcriptomic data that were available for phylogenetic analysis and to recover the overall phylogenetic signal from them. All analyses were performed on a gene-by-gene basis and also in one concatenated alignment.

METHODS

Taxon sampling and RNA sequencing

Mature spikelets from *Guadua inermis*, *Lithachne pauciflora*, *Otatea acuminata*, and *Phyllostachys aurea* were collected from tropical seasonal forests in the state of Veracruz, Mexico during the mid-dry season when florets were at full anthesis (Fig. 8). Harvested material was immediately placed in RNA-later solution (Qiagen, Valencia, CA, USA). Herbarium vouchers were collected for each species and taxonomic identities were verified (Table 9). Samples were stored at -20 °C until RNA extraction. Spikelets were homogenized in liquid nitrogen and a full RNA extraction and on-column DNase digestion was performed using the RNeasy extraction kit (Qiagen, Valecia, CA, USA) following the manufacturer's protocol. Extractions were quantified with a Nanodrop 1000 (ThermoFisher Scientific, Wilmington, DE, USA) to verify a minimum concentration of 50 ng/µl. Extractions from male and female spikelets of *L. pauciflora* were pooled to represent transcripts from both genders. All other species had bisexual flowers. Library preparation and sequencing were performed at the Carver Biotechnology Center (University of Illinois, Urbana-Champaign, IL, USA). Four samples were sequenced paired-end on the HiSeq 2000 platform (Illumina, San Diego, CA, USA). The average cDNA fragment size was 250 bp with 100 bp sequenced from each end. Sequence areas that showed a significantly low quality score ($p < 0.05$) were deleted from each set of reads using the DynamicTrim function of SolexaQA (Cox et al., 2010). Adapter sequences were trimmed from each read using CutAdapt v 1.7 (Martin, 2011).

Transcript assembly

Two transcriptome assemblies were performed for each taxon. 1) Reads were assembled de novo into contiguous sequences (contigs) using Trinity v. r20140717 (Grabherr et al., 2011). Contigs were clustered by sequence similarity using the Chrysalis function of Trinity with default settings to reduce redundancy. 2) A reference-guided assembly was performed using Tophat v. 2.0.13 (Trapnell et al., 2009) and Bowtie2 v. 0.12.7 (Langmead and Salzberg, 2012) to map reads to the previously sequenced *Phyllostachys heterocycla* nuclear scaffolds. The parameters of Bowtie2 were optimized to account for differences between the target and reference genomes by using the 'very-sensitive' setting in end-to-end mode and increasing the allowed number of mismatches in each read alignment to ten. Mapped reads were then assembled using Cufflinks v. 2.2.1 (Trapnell et al., 2012) and the gffread function of Cufflinks was used to extract transcripts from the *P. heterocycla* genome and to identify exon boundaries. Reads were mapped to these sets of transcripts using the same parameters and the consensus sequence of each was extracted using the mpileup function of samtools (Li et al., 2009). Transcripts were then clustered by 100% sequence similarity using CD-HIT v. 4.6 (Li and Godzik, 2006). All subsequent analyses were performed on both sets of assembled transcripts.

TransDecoder (http://transdecoder.sf.net) was used to predict putative reading frames, provide translations for each reading frame and further cluster transcripts. Putative reading frames were then screened for coding potential using CPAT v. 1.2.1

(Wang et al., 2013). Known coding and non-coding RNA transcripts from three well-annotated plant genomes (*A. thaliana*, *O. sativa* and *Hordeum vulgare*) were used as training data to generate a logit model for coding potential assessment. Putative reading frames were screened for a coding potential probability over 98%. Putative reading frames that fulfilled these coding potential criteria are henceforth referred to as putative coding transcripts (PCTs). PCTs were used as the basis of most transcriptomic quantitation in this study as they confidently reflect the protein coding assemblage and serve as a computationally efficient method of transcriptome annotation. TransDecoder was also used in combination with HMMER 3.0 (Finn et al, 2011) to predict functional domains in the PCTs using hidden Markov models (HMMs) and the Pfam database (Bateman et al., 2004).

Transcriptome content analysis

Both sets of PCTs were queried against the rice (*O. sativa*) proteome using BLASTX from the BLAST software package v. 2.2.25 (Altschul et al., 1997) to identify putative function. Rice was chosen as a reference because of its phylogenetic proximity to the bamboos and high level of annotation. The BLAST query was repeated twice to screen for matches to the *A. thaliana* and *P. heterocycla* proteomes. *A. thaliana* was chosen as it has the most thoroughly explored and annotated plant genome and *P. heterocycla* was chosen to identify any bamboo-specific sequences. All PCTs that matched to at least one of these proteomes will be referred to as plant-PCTs (pPCTs).

A BLASTN query of pPCTs from the de novo assembly against those produced by the reference-based assembly was performed for each taxon with a threshold e-value of 10^{-5} and an identity cutoff of 95% to determine which pPCTs were assembled by both methods and which were specific to their assembly method. All contigs were queried against the *Guadua weberbaueri* plastid genome (plastome) (GenBank: KP793062) using BLASTN with an e-value threshold of 10^{-5} to test for plastid contamination within the Illumina libraries.

MADS box identification and evolutionary analysis

MADS box homologues were identified by querying those from *B. distachyon* against our respective transcript sets using BLASTP. Homologues from *B. distachyon* were used because a thorough and recent survey of the full genome was performed to identify MADS box genes (Wei et al., 2014). Hits were filtered for redundancy and for sequences over 100 amino acids (aa) in length. Bamboo proteins were aligned conspecifically to identify copies that were assembled by both methods but differed because of assembly artifacts. Because the two most similar copies of MADS-box genes from *B. distachyon* are 93% identical, copies from the same species that shared over 95% identity were either reduced to one copy or merged to encompass regions of the protein that were determined using both assemblies. Bamboo sequences and previously identified homologues from *B. distachyon* and *O. sativa* were aligned using CLUSTALW (Thompson et al., 1994). Previously identified MADS-box genes from the genetically well-characterized *A. thaliana* were added to the alignment to aid in gene copy identification.

Geneious Pro v.8.1.7 (Biomatters, Auckland, New Zealand) was used to generate a neighbor-joining (NJ) tree for all gene copies used in this study. Sequences were assigned to clades identified in Wei et al. (2014) based on the presence of non-bamboo reference genes and were assigned names based on their analysis. Note that the copies from *B. distachyon* are labeled by the numbering system in Genbank rather than the numbering system in the tree generated by Wei et al. (2014). When the gene families of interest were identified, each set of protein sequences that corresponded to each family were grouped separately and outgroup OTUs were assigned based on the original NJ tree. A CLUSTALW alignment and NJ analysis was performed on each gene family separately to generate the best trees. All subsequent tree annotations were performed using the ETE Python package (Huerta-Cepas et al., 2010).

Nuclear orthologue phylogenetics

Only the transcripts assembled de novo were used for phylogenetic analysis, as they were less likely to be overrepresented based on preliminary results. Single-copy syntenic orthologous coding sequences, determined by Schnable et al. (2012), were extracted from the *O. sativa* and *B. distachyon* genomes. BLASTN was used to query coding sequences from *O. sativa* against pPCTs from the four bamboo species that were assembled de novo and combined with all identified coding sequences from *B. distachyon* and *P. heterocycla*. Blast hits were filtered for a maximum e-value of 10^{-5}, a minimum alignment length of 100 bp and a minimum sequence identity of 70% following methods from Zhang et al. (2014). BLASTN results were queried for instances where one copy

from *O. sativa* exhibited exactly one hit to a copy from *B. distachyon* and vice versa. The best hit from each bamboo species to each copy from *O. sativa* was located based on the highest bit-score. The homologous portions of each gene were extracted into clusters with the corresponding *O. sativa* transcript. Only groups that contained all bamboo species were used in subsequent steps. Each cluster that included five bamboo sequences and a reference sequence from *O. sativa* was then aligned using the LINSI algorithm of MAFFT (Katoh and Standley, 2013), which is optimized to align nucleotide sequences accurately.

A maximum-likelihood (ML) phylogeny was then estimated using RAxML (Stamatakis, 2006) for each alignment. The GTRGAMMA-I model was used for all trees. A majority consensus tree was produced from all ML trees using the Consense function in the PHYLIP software package (Felsenstein, 2005). An additional matrix was produced by concatenating all transcript alignments and removing all nucleotide positions that contained at least one gap to reduce ambiguity in downstream analyses. A ML analysis was also performed with RAxML using this matrix. The GTRGAMMA-I model was used and 1,000 ML bootstrap pseudoreplicates were performed with the concatenated alignment to assess topological support.

RESULTS

Transcriptomic content

A total of 214,802,213 pairs of reads were sequenced for the four taxa. A total of 505,939 unique transcripts were produced from the de novo assembly and 613,319 were produced

from the reference-based assembly. Open reading frame (ORF) detection followed by a coding potential filtering step in the de novo assembly yielded over 78,000 PCTs in *O. acuminata* and between 37,000 and 48,000 PCTs in the three remaining species. ORF detection in the reference-based assembly yielded between 43,000 and 105,000 PCTs (Table 9). The number of contigs and PCTs produced for each species and method is also reported in Table 10.

A BLASTX query of PCTs assembled de novo against the *O. sativa*, *A. thaliana*, and *P. heterocycla* proteomes (pPCTs) produced hits that covered 40–45% of PCTs in the four bamboo species. The same query of the reference-based assembly covered > 98% of PCTs in all four bamboo species (Table 10). The de novo assembly produced the highest number of unique pPCTs from the BLASTX search to the *O. sativa* proteome while the reference based assembly produced the highest number of unique pPCTs from the *P. heterocycla* proteome. All three sequence sets recovered 85.8% of the pPCTs assembled de novo and 90.7% of the pPCTs assembled using a reference genome. The numbers of pPCTs that were recovered using all reference proteomes are reported in Fig. 9.

The BLASTN comparison between the pPCTs assembled de novo and using a reference genome revealed that in *G. inermis*, *O. acuminata* and *P. aurea* 90–98% of pPCTs were represented in the de novo assembly by at least 95% sequence similarity. In *L. pauciflora* 53% of pPCTs generated using a reference genome were represented in the de novo assembly. In all taxa, 43-76% of pPCTS assembled de novo were represented in the reference-based assemblies (Fig. 10), with the *P. aurea* assembly containing the largest portion.

The HMM search on the Pfam database revealed 129,695 Pfam domains in 77,429 of the pPCTs assembled de novo (1.67 domains/ pPCT) and 379,169 domains in 248,140 of the pPCTs assembled using a reference genome (1.53 domains/ pPCT). A total of 13,650 unique domain types were represented in both species. The most numerous type of domain in both assemblies was the 'protein kinase domain.' Between 62-69% of all represented Pfam domains were present in both assemblies while 7—22% were found only in the de novo assembly and 13—25% were only found in the reference-based assembly (Fig. 11).

Querying the contig sets for similarities to the *G. weberbaueri* full plastome produced 94 matches from transcripts assembled de novo and 101 from transcripts assembled using a reference genome. The transcripts assembled de novo ranged from 1,286-47,612 bp and those assembled using a reference ranged from 352-6,151 bp.

Exploration of expressed floral genes

In the following, the terms "genes" and "gene copies" refer to those that were expressed and assembled from our transcriptomic data set. After elimination of redundant copies from the transcript sets, a total of 72 MADS-box genes were identified within the six gene families of interest: A, B, C/D, E-class, *SOC-like* and *SVP-like* MADS-box genes. The fewest number of copies were identified in the *SVP-like* gene family (4) and the most were identified in the B-class gene family (17). The fewest number of identified copies were expressed in *G. inermis* (13), while the other three taxa expressed comparable

numbers (19-20). The total number of copies expressed by each taxon for all gene families is reported in Table 11. Bamboo gene copies that exhibit putative orthology to copies from *O. sativa* are listed in Table 12.

Six neighbor-joining gene trees were generated using peptides from each of the six gene families. The gene tree topologies did not uniformly reflect previously recovered taxonomic relationships among species, but did exhibit notable patterns. A sister relationship to other bamboo species was found in 25% of all copies from *L. pauciflora*, with the other 75% associating either with copies from reference taxa or with clades composed of both reference and bamboo copies. A total of 55% of copies from *P. aurea*, 69.2% from *G. inermis*, and 73.7% from *O. acuminata* exhibited a sister relationship to other bamboo species.

The relationships between copies from bamboos and the reference taxa from other grass subfamilies were noted when subtrees contained representative copies from Bambusoideae, Oryzoideae (*O. sativa*) and Pooideae (*B. distachyon*) and the copies from Bambusoideae exhibited monophyly. These subtrees could be classified into three distinct categories: 1) Bamboo copies exhibit a sister relationship to copies from *O. sativa* + *B. distachyon*, 2) the copy from *B. distachyon* exhibits monophyly with bamboos and this clade exhibits a sister relationship to copies from *O. sativa*, 3) the copy from *O. sativa* exhibits monophyly with bamboos and this clade exhibits a sister relationship to copies from *B. distachyon* (see clades marked with asterisks in Figs. 12–14). A total of eighteen distinct instances could be classified into these categories. Eight instances fell

under category 1, while five instances could be classified under each of categories 2 and 3.

The A-class genes formed two distinct clades that included only grasses and one additional large clade that included grasses and *A. thaliana* (Fig. 12-A). Clade **a** associates with *OsMADS18* and includes one copy from *L. pauciflora, O. acuminata* and *P. aurea*, as well as a gene copy from *B. distachyon.* The second grass-specific clade (**b**) consisted of genes from only two species of grasses (*OsMADS20 + BdMADS20*) and did not include any bamboo copies. Another grass clade associated with the *FRUITFULL*, *CAULIFLOWER*, and *APETALA1* genes from *A. thaliana* (**c**). Clade **c** included two gene copies from all bamboo species except for *O. acuminata,* which was represented by one copy.

The B-class genes (Fig. 13-B) formed four clades that associate with copies of MADS-box genes from *A. thaliana* and a fifth grass-specific clade. Clade **a**, which associated with *AGL32* and *GORDITA* from *A. thaliana,* united a bamboo specific subclade with one copy each from *O. sativa* and *B. distachyon*. One additional gene copy from both *P. aurea* and *L. pauciflora* are also included in clade **a** diverging in intermediate positions. Clade **b**, which associated with *AGL12/XAL* from *A. thaliana*, comprised one copy from *O. acuminata* and two copies from *L. pauciflora.* Each of the copies from *L. pauciflora* in clade **b** formed a sister relationship to a copy from *O. sativa* or *B. distachyon*. Clade **c** associated with *APETALA1* from *A. thaliana* and contained one copy each from *G. inermis, L. pauciflora* and *P. aurea*. Clade **d** associated with *PISTILLATA* from *A. thaliana* and contained two copies from *O. acuminata*, and one

from each of the other three bamboo species. Clade **d** also possessed two copies each from *O. sativa* and *B. distachyon*. The grass-specific clade (**e**) comprised one copy from *L. pauciflora*, which associated with copies *OsMADS29* and *BdMADS29*.

The C/D class genes are paraphyletic and were combined into one tree (Fig 12-C/D). The D-class genes from grasses form clade **a**, while another clade (**b**) includes C-class genes from grasses (subclades **c** and **d**), C-class genes from *A. thaliana* (**e**) and one D-class gene from *A. thaliana* (*SEEDSTICK*). The four copies from *A. thaliana* formed a clade with one copy from *O. acuminata* and no other grasses (**e + f**). The remainder of the clades (**a**, **c**, **d**) were grass specific. Clade **a-1**, which associates with *OsMADS21* contains two copies from *P. aurea*, one copy from *G. inermis* and one gene copy from *L. pauciflora*. The other exclusively C-class clade (**a-2**) associates with *OsMADS13* and contains one gene copy from all bamboo species except for *G. inermis*. The next grass-specific clade of D-class genes (**c**) associated with *OsMADS3* and *BdMADS3*, contained one gene copy from all four bamboo species. The other grass-specific D-class clade (**d**) contained one gene copy from *O. acuminata*, one from *P. aurea*, two copies from *O. sativa* and one from *B. distachyon*.

The E-class gene tree (Fig 13-E) formed three clades that associated with MADS-box gene copies from *A. thaliana* (**a, b, c**) and one grass specific clade (**d**). One gene copy from *O. acuminata* and one from *P. aurea* associated with *RSB* from *A. thaliana* in clade **a** along with *OsMADS6*, *BdMADS6*, and *OsMADS17*. Clade **b** contained *SEPALLATA3* from *A. thaliana*, which formed a sister relationship to two subclades of grasses that each associated with a gene copy from *O. sativa*. The first subclade (**b-1**)

contained one gene copy from *G. inermis*, one copy from *L. pauciflora,* and *OsMADS8.* The second subclade (**b-2**) contained one gene copy from *O. acuminata,* one copy from *P. aurea, OsMADS7* and *BdMADS7*. Clade **c** contained one gene copy from *L. pauciflora* and one from *O. acuminata*, which formed a sister relationship to *SEPALLATA1+SEPALLATA2* from *A. thaliana*. The distribution of gene copies from *O. sativa* and *B. distachyon* suggests that the grass specific clade (**d**) diverged into three subclades. The first subclade (**d-1**) contained one gene copy from *P. aurea* and associates with *OsMADS34* and *BdMADS34*. The second subclade (**d-2**) contained one gene copy from all species but *L. pauciflora* and associates with *OsMADS55* and *BdMADS55*. The third subclade (**d-3**) contains gene copies from all species except for *P. aurea* and associates with *OsMADS1* and *BdMADS1*.

The *SOC-Like* gene tree (Fig 14-SOC) formed a clade of grass specific genes (**a**) separate from five *A. thaliana* copies. One subclade of grasses (**b**) contained copies from all four species of bamboos, but also formed a sister relationship to a second gene copy from *L. pauciflora.* Subclade **b** also associated with *OsMADS56, BdMADS56* and *BdMADS50*. This subclade (**c**) contained three copies from *P. aurea*, two from *O. acuminata* and one each from *G. inermis* and *L. pauciflora.* The second subclade (**c**) is sparsely populated with bamboo copies as four copies from *B. distachyon* (*BdMADS56, BdMADS50, BdMADS7, BdMADS22*) are present along with two copies from *O. sativa* (*OsMADS56, OsMADS37*), but only seven bamboo copies, from four species, are present. The *SVP-like* gene tree (Fig 14-SVP) contains only one copy of each bamboo species with the woody bamboos forming clade **a**. The copy from *L. pauciflora*, which is

separated from the woody bamboo clade, formed a sister relationship to a copy from *O. sativa.*

Nuclear orthologue phylogenetics

A total of 3,878 clusters were produced that met the criteria set for this study. The length of each alignment ranged from 219--10,755 bp. The concatenated alignment had a length of 5,736,540 bp, which was reduced to 2,698,410 bp after gapped positions were removed.

The concatenated analysis produced a fully-resolved ML tree and was rooted at *O. sativa* based on previous phylogenetic results (Fig. 15). In this case, *B. distachyon* formed a sister relationship with the rest of the taxa (bamboos). Within Bambusoideae, *L. pauciflora* formed a sister relationship with the rest of the bamboos, which were all woody. The tropical woody bamboos, *G. inermis* + *O. acuminata* formed a sister relationship to the temperate woody bamboos, *P. aurea* + *P. heterocycla*. All relationships were supported at 100% ML bootstrap.

Out of the 3,878 best gene trees, 2,374 trees (61.21%) produced a monophyletic Bambusoideae when rooted at *O. sativa*. Monophyly in the woody bamboos was recovered in 2,412 trees (62.19%). A *P. aurea* + *P. heterocycla* sister relationship was recovered in 2,956 trees (76.22%) and a *G. inermis* + *O. acuminata* sister relationship was recovered in 3,339 trees (86.10%). A relationship consistent with the chloroplast phylogeny was recovered in 215 trees (5.54%).

DISCUSSION

Comparison of de novo and reference-based transcriptome assemblies

While both de novo and reference-based assemblies have been used to describe full transcriptomes, these two methods, which were performed here on identical sets of reads, exhibited strikingly different transcriptomic results. The reference-based assembly produced PCT sets that were consistently shown to have higher percentages of pPCTs. This clearly reflects the nature of reference-based methods in which only reads that met a sequence similarity threshold to a previously-sequenced plant genome were assembled. All species used here exhibited either higher or comparable levels of pPCT abundance in the reference-based assembly.

The pPCTs recovered using reference proteomes from *A. thaliana*, *O. sativa* and *P. heterocycla*, could be placed into general overlapping subgroups. Those pPCTs recovered using *A. thaliana* were representative of the transcripts that could be generally found in most angiosperms given the phylogenetic distance of the eudicots to the bamboos. Those recovered using *O. sativa* were representative of genes that could be found in grasses, and those recovered using *P. heterocycla* represented genes specific to bamboos. While these taxonomic levels are not precisely defined (i.e. some hits to the *O. sativa* proteome may be indicative of BOP-clade specific transcripts), the overlap exhibits predictable patterns. In both assemblies, pPCTs that were uniquely recovered using *A. thaliana* form the smallest group while the largest group uniquely recovered in the reference assembly is from *P. heterocycla*. However, more pPCTs were uniquely recovered from *O. sativa* in the de novo assembly. This could be indicative of a

weakness in the de novo assembly in which transcripts are present, but not at high enough abundances to assemble them without the aid of a reference genome. It also could be an artifact of the reference-based assembly and reliance on the *P. heterocycla* genome.

When a 95% sequence identity threshold was used to assess redundancy, a much larger portion of the pPCTs assembled de novo is unique to each assembly, except those from *L. pauciflora*. For the three woody species, this would indicate that the de novo assembly produced a set of unique transcripts while the reference based assembly produced transcripts that were largely represented in the de novo assembly. This is likely indicative of under-assembly in the reference-based trancripts. While this could be an indication that a reference-based assembly is largely unnecessary and could artificially inflate the number of transcripts produced, the number of uniquely represented transcripts in the reference-based *L. pauciflora* assembly may suggest otherwise.

The reference-based assembly likely recovered a large number of transcripts in *L. pauciflora* with coverage too low to be assembled de novo. This does not reflect any inherent property of the *L. pauciflora* reads (lower number, quality, etc.) but may reflect a difference in genomic properties. While exact ploidy levels are not known for the species analyzed in this study, herbaceous bamboos are known to be primarily diploid, while woody bamboos tend to exhibit higher levels of ploidy likely resulting from hybridizations (Hunziker et al., 1982; Chokthaweepanich et al. 2014; Triplett et al., 2014). A lower ploidy level in *L. pauciflora* could have been conducive to a larger spread of transcriptomic representation due to higher coverage levels for each gene.

Further support for performing both types of assembly comes from the Pfam analysis in the pPCTs. Only 62-69% of all represented Pfam domains were present in both assemblies even though both shared a considerable proportion of unique protein domains (Fig. 11). The reference-based assembly for *P. aurea* produced more than three times the number of unique Pfam domains than the de novo assembly. While the other three produced fewer, this is most likely due to the phylogenetic proximity of *P. aurea* to the congeneric reference genome (Triplett and Clark, 2010).

Presence of plastid genome sequences

Although the Illumina libraries were sequenced after selection for transcripts containing poly-A tails, sequences exhibiting plastome homology were likely assembled due to the high number of plastids present in each plant cell. The presence of plastid sequences may have been retained during the poly-A selection step due to the AT richness of the plastome. A large number of these transcripts, which were assembled de novo, were much larger than mRNA transcripts that were observed in eukaryotic plastomes > 10 kbp and contained more than one coding sequence. One explanation for these unusually large transcripts is that plastid genomes are typically very compact and contain relatively small intergenic regions. If overlapping UTRs were present in the sequenced transcripts, it could be an artifact of the de novo assembly method, which takes no reference genome into account and produces assemblies based solely on sequence identity.

Floral gene analysis

Because the field-harvested sources of our RNA extracts were spikelets with fully developed and emergent florets, not all MADS-box genes (A-E class) were expected to be expressed. This is apparent as significantly fewer MADS box genes were found in each bamboo transcriptome (13-20) than the 57 that were found in the fully sequenced *B. distachyon* genome (Wei et al., 2014). Between two and five genes were identified in all classes for each taxon except for *L. pauciflora*, which expressed seven distinct B-class genes. Because B-class genes are important in lodicule and stamen development in grasses (Whipple et al., 2004), they are required in the development of both types of unisexual florets that are produced by herbaceous bamboos. We hypothesize that duplications of two B class genes may have allowed separate copies to be expressed differentially in either male (lodicule and stamen development) or female (lodicule only development) florets. Though *O. sativa* and *B. distachyon* have bisexual florets and are both represented with seven copies, their data is genomic and the lower number of transcripts in the three woody bamboos is likely a result of differential expression. The first of these, involving gene copies B1 and B2, are both found in clade **a** (Fig. 13B). Gene copy B1 clusters with (*OsMADS31* + *BdMADS31*). Gene copy B2 is sister to all of the grass-specific members of this clade, rather than to any specific gene copy, although copy B1 of *P. aurea* is the next copy in the clade to diverge. In the second case there are two copies, B3 and B4, which are found in clade **b** (Fig. 13B) and cluster with *OsMADS26* and *BdMADS26*, respectively. Unexpectedly copy B5 from *L. pauciflora* does not cluster with any other copies from Bambusoideae, so its origin is obscure. This

hypothesis could be verified by sequencing B-class genes from separate staminate and carpellate florets.

In several cases, correlations between gene copy number and differences in floral phenotypes can be hypothesized. The *PISTILLATA* homologues from *O. sativa*, *OsMADS2* and *OsMADS4* were shown to have complementary importance in lodicule development and about equal importance in stamen development (Yao et al., 2008). One copy each from *O. acuminata* and *P. aurea* associate with *OsMADS2* (Fig. 13-B; clade d), which may be indicative of phenotype as both have three stamens compared to *G. inermis*, which has six (Judziewicz et al., 1999). However, sister to the *OsMADS2-4* subclade within clade d are one copy each from *G. inermis* and *O. acuminata* which may be indicative of a duplication resulting in a novel, bamboo-specific, B-class gene. Another potentially bamboo-specific gene can be found in one copy from *L. pauciflora* and *O. acuminata* to the E-class *SEPALLATA1-2* genes from *A. thaliana* (Fig. 13-E; clade c). The placement of these copies seems to be indicative of a gene duplication in bamboos or a copy deletion in non-bamboo grasses.

Within the C/D-class clade, which are paraphyletic in regard to function as previously noted in broad studies of fruit development genes (Pabón-Mora et al. 2014), *OsMADS3* and *OsMADS58* are known to be C-class genes. *OsMADS3* has been shown to be important in carpel diversification (Yamaguchi et al., 2004) and clusters with copies from *G. inermis*, *O. acuminata*, and *L. pauciflora* (Fig. 12-C/D; clade c). These three genes also distantly associate to a copy from *P. aurea*. This could potentially be indicative of the phenotypic difference in carpels between the clade formed by copies

from *G. inermis, O. acuminata*. *L. pauciflora, O. sativa,* and B. *distachyon*, which have two stigmas, and *P. aurea*, which has three stigmas and forms a sister relationship with the two-stigma clade.

Another potential connection to phenotype can be found in the gene tree for *SVP-like* genes, which has been tied to flowering time. One copy from each species of bamboo was present in this tree (Fig 14-*SVP*). The copy from *L. pauciflora* is the immediate sister to a copy from *O. sativa*, which also flowers annually, rather than with the other three bamboo species, the latter of which formed clade **a** and exhibited 92.6% sequence similarity. The copy from *L. pauciflora* exhibited between 69-82% sequence similarity to the other bamboo copies. This may be explained by the different phenological patterns found among these species; *L. pauciflora* is a perennial that flowers approximately annually while the other three taxa flower at very long intervals (Clark et al., 2015).

These phenotypic connections to expressed gene copy number and evolutionary history are interesting, and could be confirmed with subsequent testing. *In situ* hybridization and transcriptome sampling at different stages of floral development could be performed to verify these hypotheses. The bioinformatics-based survey performed in this study is a foundation to further elucidating the complete flowering mechanisms of these, and other, bamboo species.

The presence of sister orthologues of *O. sativa* and *B. distachyon* that do not associate with bamboo copies is observed at least three times (Fig. 12-A, clade b: *OsMADS20+BdMADS20*; Fig. 13-B, clade b: *OsMADS33-BdMADS33;* clade e:

OsMADS30-Bd_ZMM17-Like). The most probable explanation for this pattern is that orthologues (or close homologues) of these genes were not expressed in the four tissues that were used in this study. This is especially likely as some A/B-class genes are known to be expressed earlier in floral development and the florets harvested for this study were fully developed. The second possibility is that there was a deletion of these copies in the Bambusoideae. The possibility of gene duplication in the *B. distachyon* and *O. sativa* lineage(s) is very unlikely as previous phylogenetic studies have placed the Bambusoideae either sister to *B. distachyon* or sister to *O. sativa* (see below).

One important caveat to any of the transcriptome comparisons made within this study is that inconsistencies may arise from the method of tissue retrieval and the study system used. While the floral tissue was harvested from spikelets of approximately equal maturity, the stresses and conditions endured by each plant (i.e. soil type, climate, herbivory) may have been significantly different. This method of collecting floral tissue is necessary when using bamboos as a study system since their flowering cycles are typically very long and unpredictable, and they are difficult to cultivate as flowering specimens under greenhouse conditions with few exceptions (Lin et al., 2005). Genome sequencing followed by an extensive survey for functional genes would allow us to more confidently confirm the presence or absence of specific gene copies.

Floral genes and phylogeny

The repeated instances of gene copies from *L. pauciflora* being isolated from copies from other bamboos in our gene trees (Figs. 12--14) could be explained by the differences between Olyreae and the two woody bamboo tribes (Arundinarieae and Bambuseae) in

phenology, sexual systems, floral development and structure. This could be an indication that Olyreae are evolutionarily separate from the woody bamboo lineages, but contradicts previous studies that have placed Olyreae in a sister relationship to Bambuseae (Wu and Ge, 2012; Kelchner et al., 2013; Wysocki et al., 2015). This potential indication of shared ancestry between Arundinarieae and Bambuseae is compatible with the results of Triplett et al. (2014), who proposed a monophyletic woody bamboo clade (Bambuseae + Arundinarieae) on the basis of phylogenetic analysis of single-copy nuclear markers. However, complete genomic sequencing and annotation would be required to rule out the possibility that these topologies are a result of unexpressed, and therefore missing, copies.

Most of the subfamilial grass phylogeny that could be inferred from these trees showed a topology in which bamboos are more closely related to *O. sativa* than to *B. distachyon*. This contrasts markedly with previously published phylogenetic studies that use plastid markers (Wu and Ge, 2012; Kelchner et al., 2013; Wysocki et al., 2015), which place the Oryzoideae (*O. sativa*) in a sister relationship to the Bambusoideae + Pooideae (*B. distachyon*) clade. However, some studies that used plastid markers have placed Bambusoideae in a sister relationship to Oryzoideae (Clark et al, 1995; GPWG, 2001).

The use of nuclear transcripts as phylogenetic markers has been controversial because of the shorter length of single transcripts, selective effects on coding mRNAs and the ambiguity in differentiating orthologues from paralogues (or other homologues). The aforementioned speculations based on MADS box gene tree topological patterns can be

tested extensively, using a large variety of nuclear markers, to draw robust conclusions about the evolutionary relationships among these taxa.

Nuclear orthologue phylogenetics

The recovery of two conflicting tribal topologies in the Bambusoideae brings two main questions into consideration: 1) Which topology reflects the actual species tree and evolutionary history of Bambusoideae? 2) Which evolutionary events would cause these conflicts to emerge? The trees produced using concatenated nuclear genes strongly supported the hypothesis of a woody bamboo clade. A monophyletic relationship among the woody bamboos is also supported in the majority of separate gene trees. We will refer to this topology, ((Arundinarieae+Bambuseae), Olyreae), as the 'nuclear hypothesis' of bamboo evolution. Phylogenetic trees that use plastid genomes have recovered a well-supported paraphyly in the woody bamboos, ((Olyreae+Bambuseae), Arundinarieae), which we will refer to as the 'plastid hypothesis.'

Based purely on the quantity of data, the nuclear hypothesis is supported by over an order of magnitude more nucleotides than the plastid hypothesis. The nuclear hypothesis is supported by the phylogenetic signal given by over 3,700 loci, while the complete plastid genome is inherited cytoplasmically and theoretically gives the phylogenetic signal equivalent to one locus. However, the single-locus attribute of the plastid genome gives it a much higher degree of certainty in its use as a phylogenetic marker between plant species. Although care was taken to minimize our level of uncertainty, nuclear genes can have multiple paralogous copies. Their use as

phylogenetic markers can also be complicated by allopolyploidy, which many grasses, including woody bamboos, have been shown to exhibit (Levy and Feldman, 2002).

Although the use of morphological characteristics in the determination of phylogenetic relatedness is controversial, the morphological and phenological characteristics of the three bamboo tribes are notable. Arundinarieae and Bambuseae share a suite of characteristics with highly lignified culms, bisexual flowers and gregarious monocarpy. Olyreae have shoots with significantly less lignification, unisexual florets and annual flowering. The nuclear hypothesis suggests a single origin of these characteristics and the plastid hypothesis suggests two origins or one origin followed by a loss of these characteristics in Olyreae. The duplicate origin of similar characteristics seems unlikely, but a loss of certain characters could be biologically feasible.

A hypothesis for the validity of both phylogenetic signals involves an ancient hybridization. With the species tree following the nuclear hypothesis, the Bambuseae and Arundinarieae would exhibit a sister relationship to Olyreae. If a progenitor species of Bambuseae had hybridized with a sympatric progenitor species of Olyreae, followed by a back-cross in the paternal species, the maternally inherited plastid signal would place Bambuseae phylogenetically sister to Olyreae rather than Arundinarieae. An alternative explanation for the incongruence of phylogenetic signal could be selection in which either tropical bamboos (Bambuseae and Olyreae) or woody bamboos (Bambuseae and Arundinarieae) accumulated homoplasious mutations. However, it would be very unlikely that over 60% of the genes sampled in our analyses skewed the phylogenetic

signal identically, placing Bambuseae sister to Arundinarieae, due to selection. Long-branch attraction can be eliminated as a possibility because both Bambuseae and Arundinarieae produce very short branches within each tribe (Wysocki et al., 2015).

Triplett et al. (2014) also recovered monophyly in the woody bamboos using three low copy nuclear genes. However, each gene copy was classified into respective ancient genomes based on the hypothesis of woody bamboos being a product of allopolyploidization. Our study assumed orthology based on the highest sequence similarity and the presence of one orthologous copy per species. While the single-orthologue approach does not account for multiple orthologous copies from allopolyploids, the overall phylogenetic signal from the concatenated alignment and trees is robust. If the putative orthologues from each taxon originated from different progenitor woody bamboo genomes, we might expect the support for nodes within the woody bamboo clade to reflect this.

Table 9. Species used in this study, with collection sites, collectors, collector numbers, the herbarium where the specimen vouchers are deposited and the number of paired-end reads generated for each specimen.

Taxon	**Country/State**	**Collectors/Number**	**Herbarium**	Number of read pairs
***Guadua inermis* Rupr. ex E. Fourn**	Mexico, Veracruz	E. Ruiz-Sanchez & W. Wysocki, 466	IEB	54,488,684
***Otatea acuminata* (Munro) Calderón & Soderstr.**	Mexico, Veracruz	E. Ruiz-Sanchez & W. Wysocki, 469	IEB	55,890,954
***Phyllostachys aurea* Carrière ex Rivière & C. Rivière**	Mexico, Veracruz	E. Ruiz-Sanchez & W. Wysocki, 470	IEB	46,405,022
***Lithachne pauciflora* (Sw.) P. Beauv.**	Mexico, Veracruz	E. Ruiz-Sanchez & W. Wysocki, 470a	IEB	58,017,553

Table 10. The number of contigs, PCTs and pPCTs generated for each taxon. Results from both assemblies are reported here.

	Taxon	Contigs	PCTS	pPCTs
De novo	*G. inermis*	108438	44252	19254
	O. acuminata	111831	78443	31498
	P. aurea	88256	48001	21033
	L. pauciflora	197414	37848	17212
Reference	*G. inermis*	96762	43226	42788
	O. acuminata	180171	97094	95665
	P. aurea	197627	104413	103165
	L. pauciflora	144759	77433	76278

Table 11. A total of 72 MADS-box genes were identified in this study. Numbers from each gene class and taxon are reported.

Class	G. inermis	O. acuminata	P. aurea	L. pauciflora	*Total*
A	2	2	3	3	10
B	3	4	3	7	17
C/D	2	4	5	3	14
E	3	5	4	3	15
SOC-Like	2	3	4	3	12
SVP-Like	1	1	1	1	4
Total	13	19	20	20	72

Table 12. MADS-box gene copies from bamboos matched to corresponding orthlogs from *Oryza sativa*. MADS-box gene classes are noted. Gene copies in paretheses denote orthlogs that were inferred ambiguously based on topological relationships.

Gene Class	***Oryza sativa***	***Guadua inermis***	***Otatea acuminata***	***Phyllostachys aurea***	***Lithachne pauciflora***
A	*OsMADS14*	Guadua_MADS_A_2	-	Phyllostachys_MADS_A_3	(Lithachne_MADS_A_3)
	OsMADS15	-	-	Phyllostachys_MADS_A_2	Lithachne_MADS_A_2
	OsMADS18	-	Otatea_MADS_A_1	Phyllostachys_MADS_A_1	Lithachne_MADS_A_1
	OsMADS20	-	-	-	-
B	*OsMADS2*	-	Otatea_MADS_B_4	Phyllostachys_MADS_B_3	-
	OsMADS4	-	-	-	Lithachne_MADS_B_7
	OsMADS16	Guadua_MADS_B_2	-	Phyllostachys_MADS_B_2	Lithachne_MADS_B_6
	OsMADS26	-	Otatea_MADS_B_2	-	Lithachne_MADS_B_3,4
	OsMADS29	-	-	-	Lithachne_MADS_B_5
	OsMADS30	-	-	-	-
	OsMADS31	Guadua_MADS_B_1	Otatea_MADS_B_1	Phyllostachys_MADS_B_1	Lithachne_MADS_B_1,2
	OsMADS33	-	-	-	-
C/D	*OsMADS3*	Guadua_MADS_C/D_2	Otatea_MADS_C/D_3	Phyllostachys_MADS_C/D_4	Lithachne_MADS_C/D_3

	OsMADS13	-	Otatea_MADS_C/D_2	Phyllostachys_MADS_C/D_3	Lithachne_MADS_C/D_2
	OsMADS21	Guadua_MADS_C/D_1	-	Phyllostachys_MADS_C/D_1,2	Lithachne_MADS_C/D_1
	OsMADS58	-	(Otatea_MADS_C/D_4)	(Phyllostachys_MADS_C/D_5)	-
	OsMADS66	-	(Otatea_MADS_C/D_4)	(Phyllostachys_MADS_C/D_5)	-
E	*OsMADS1*	Guadua_MADS_E_3	Otatea_MADS_E_1	-	Lithachne_MADS_E_2
	OsMADS5	Guadua_MADS_E_2	Otatea_MADS_E_3	Phyllostachys_MADS_E_4	-
	OsMADS6	-	Otatea_MADS_E_2	Phyllostachys_MADS_E_2	-
	OsMADS7	-	Otatea_MADS_E_1	Phyllostachys_MADS_E_1	-
	OsMADS8	Guadua_MADS_E_1	-	-	Lithachne_MADS_E_1
	OsMADS17	-	-	-	-
	OsMADS34	-	-	Phyllostachys_MADS_E_3	-
SOC	*OsMADS37*	-	Otatea_SOC-Like_3	Phyllostachys_SOC-Like_3	-
	OsMADS56	Guadua_SOC-Like_1	Otatea_SOC-Like_1	Phyllostachys_SOC-Like_1	Lithachne_SOC-Like_1,2
	OsMADS65	Guadua_SOC-Like_2	Otatea_SOC-Like_2	Phyllostachys_SOC-Like_2	Lithachne_SOC-Like_3
SVP	*OsMADS22*	Guadua_SVP-Like_1	Otatea_SVP-Like_1	Phyllostachys_SVP-Like_1	-
	OsMADS47	-	-	-	-
	OsMADS55	-	-	-	Lithachne_SVP-Like_1

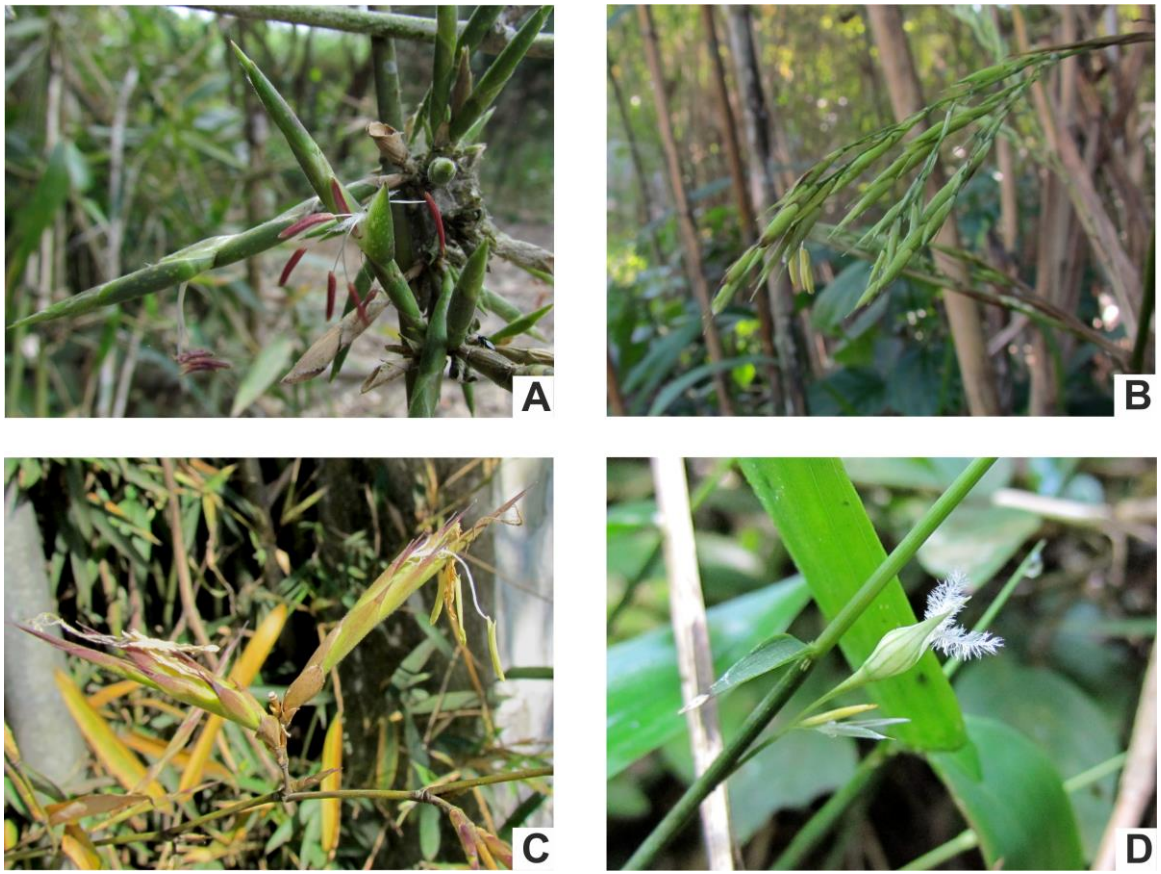


Figure 8. A. *Guadua inermis* pseudospikelets. B. *Otatea acuminata*, spikelets. C. *Phyllostachys aurea*, pseudospikelets. D. *Lithachne pauciflora*, male and female spikelets (floral structures of the other three species are bisexual). Note that photos A, B and C represent the actual specimens collected for this study. Photo D represents a separate individual of the same species. Photos by E. Ruiz-Sanchez.

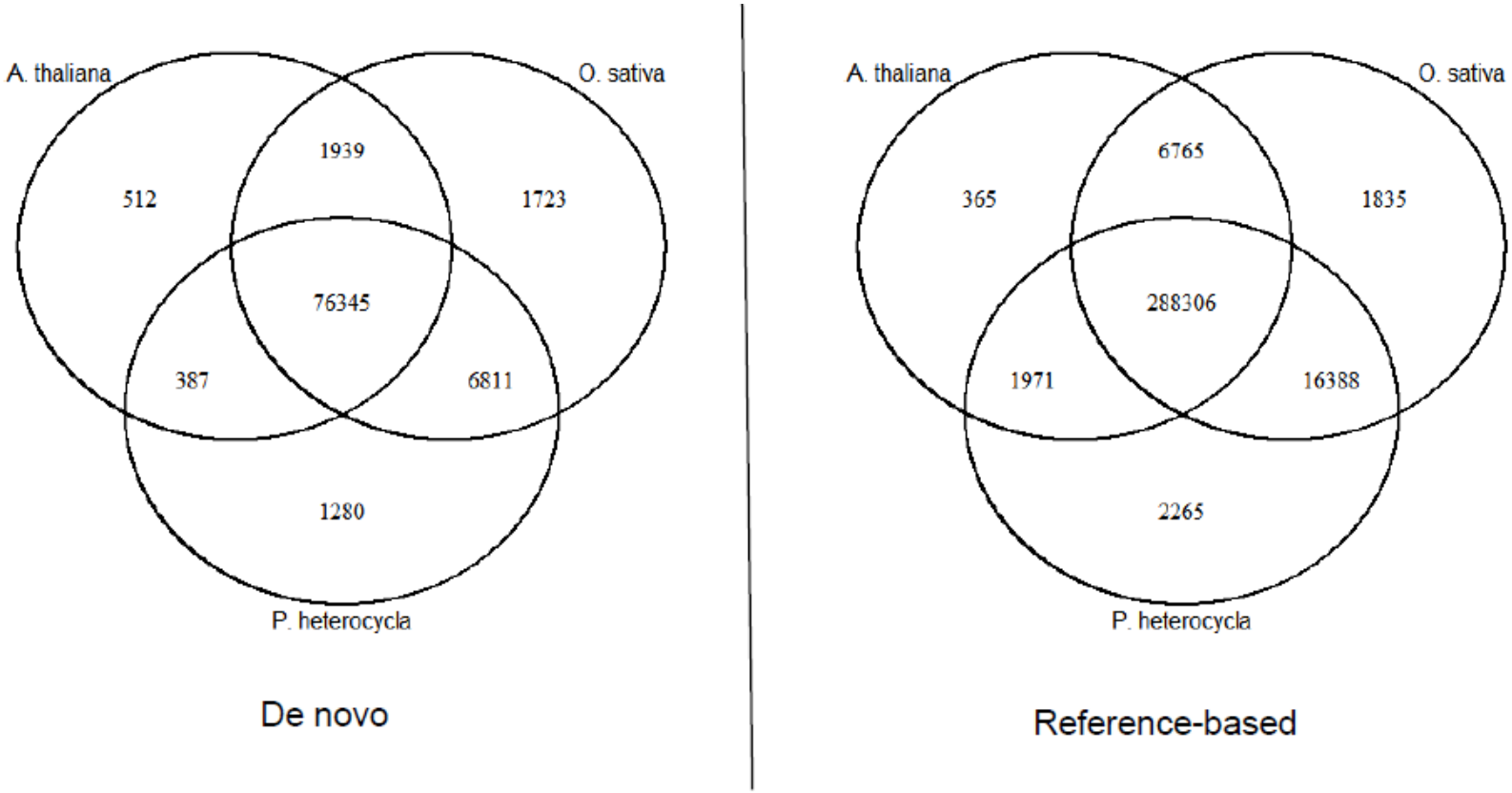


Figure 9. Venn diagrams reporting the number of pPCT hits that are unique to each reference proteome and shared by them. Diagrams are separated by assembly type. Venn diagrams were generated using the VennDiagram R-package.

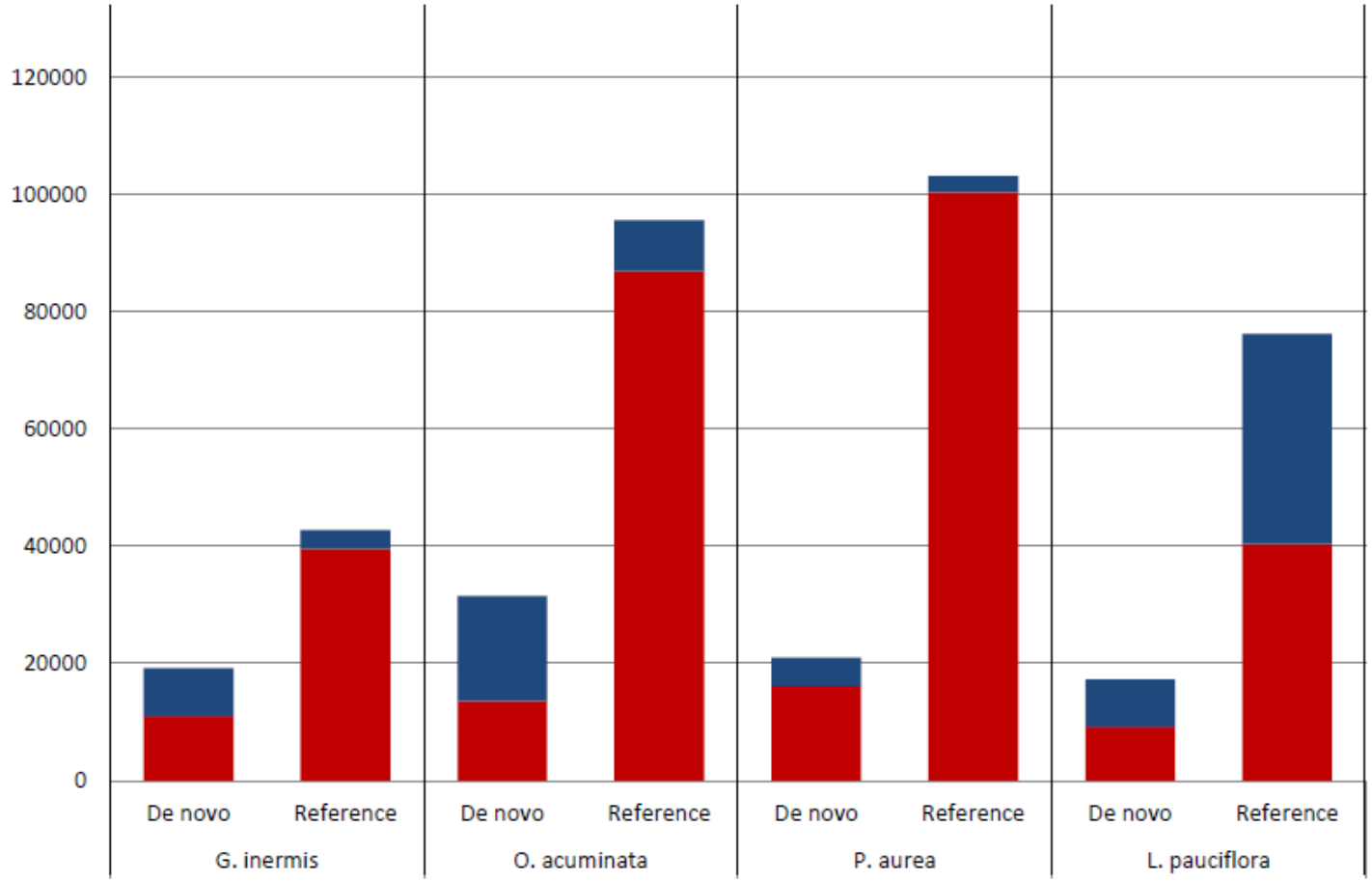


Figure 10. Bar graph indicating the number of pPCTs (putative transcripts that show a close plant homolog) for each taxon for both assemblies. The red portion of each bar indicates the number of redundant transcripts that exhibit at least 95% nucleotide sequence identity to the other assembly. The blue portion represents transcripts that do not reach this criterion and are unique to their respective assembly.

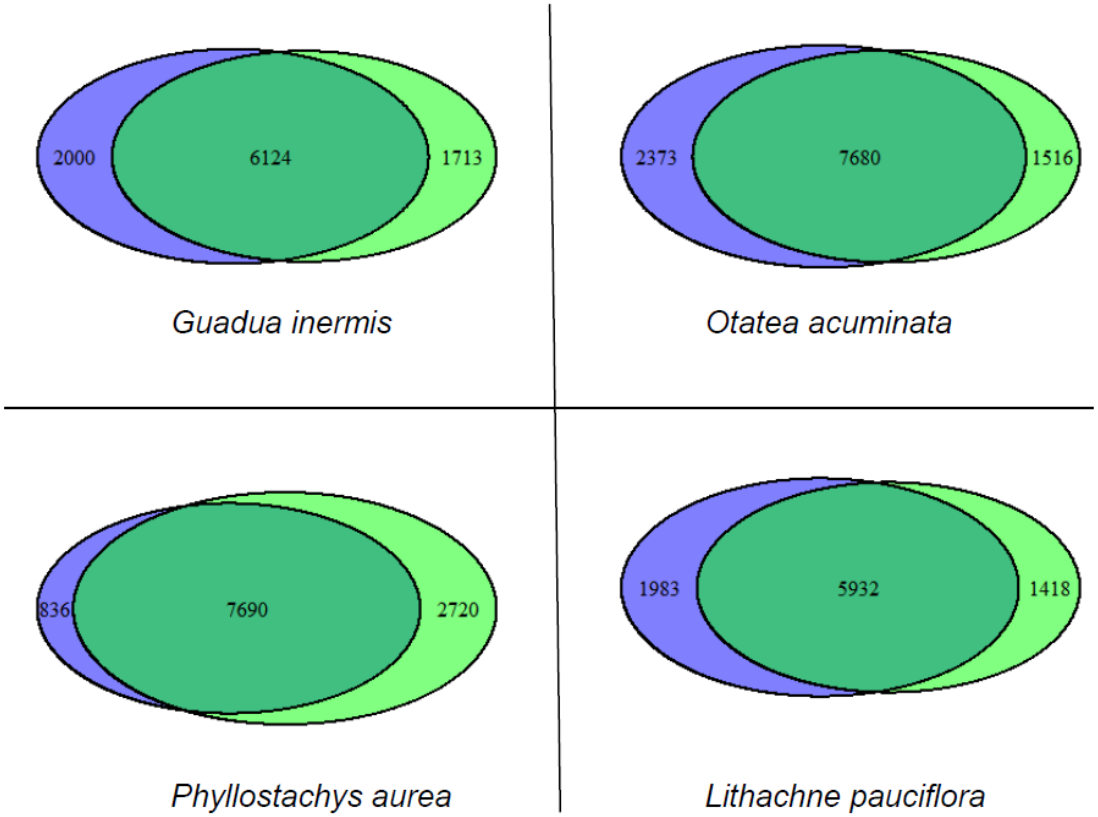
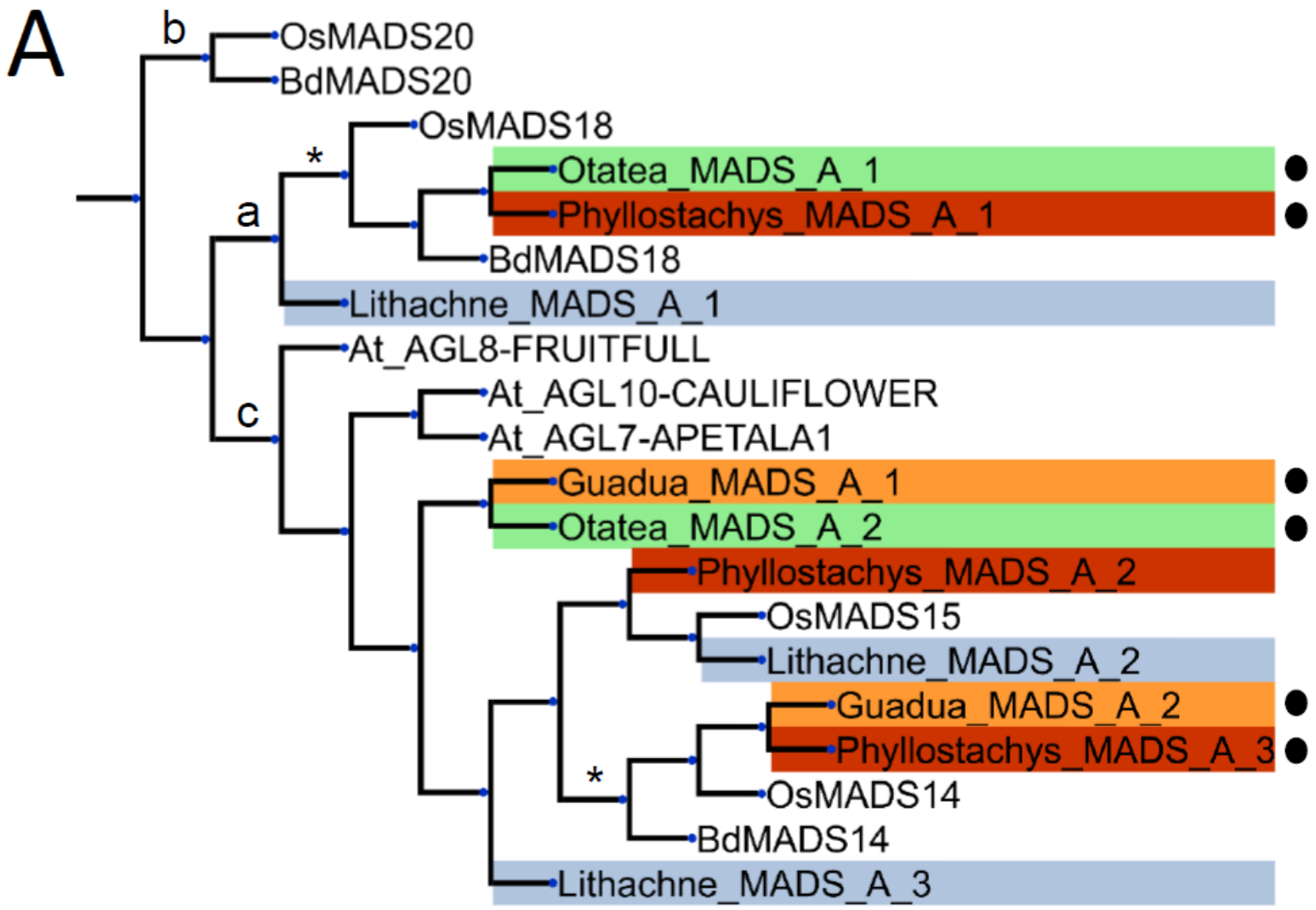


Figure 11. Venn diagrams reporting the number of Pfam domains that are unique to each assembly and shared by both assemblies. Diagrams are separated by taxon. Venn diagrams are proportional to their values and were generated using the VennDiagram R-package.


A
b
OsMADS20
BdMADS20
OsMADS18
*
Otatea_MADS_A_1
Phyllostachys_MADS_A_1
a
BdMADS18
Lithachne_MADS_A_1
At_AGL8-FRUITFULL
At_AGL10-CAULIFLOWER
c
At_AGL7-APETALA1
Guadua_MADS_A_1
Otatea_MADS_A_2
Phyllostachys_MADS_A_2
OsMADS15
Lithachne_MADS_A_2
Guadua_MADS_A_2
Phyllostachys_MADS_A_3
*
OsMADS14
BdMADS14
Lithachne_MADS_A_3

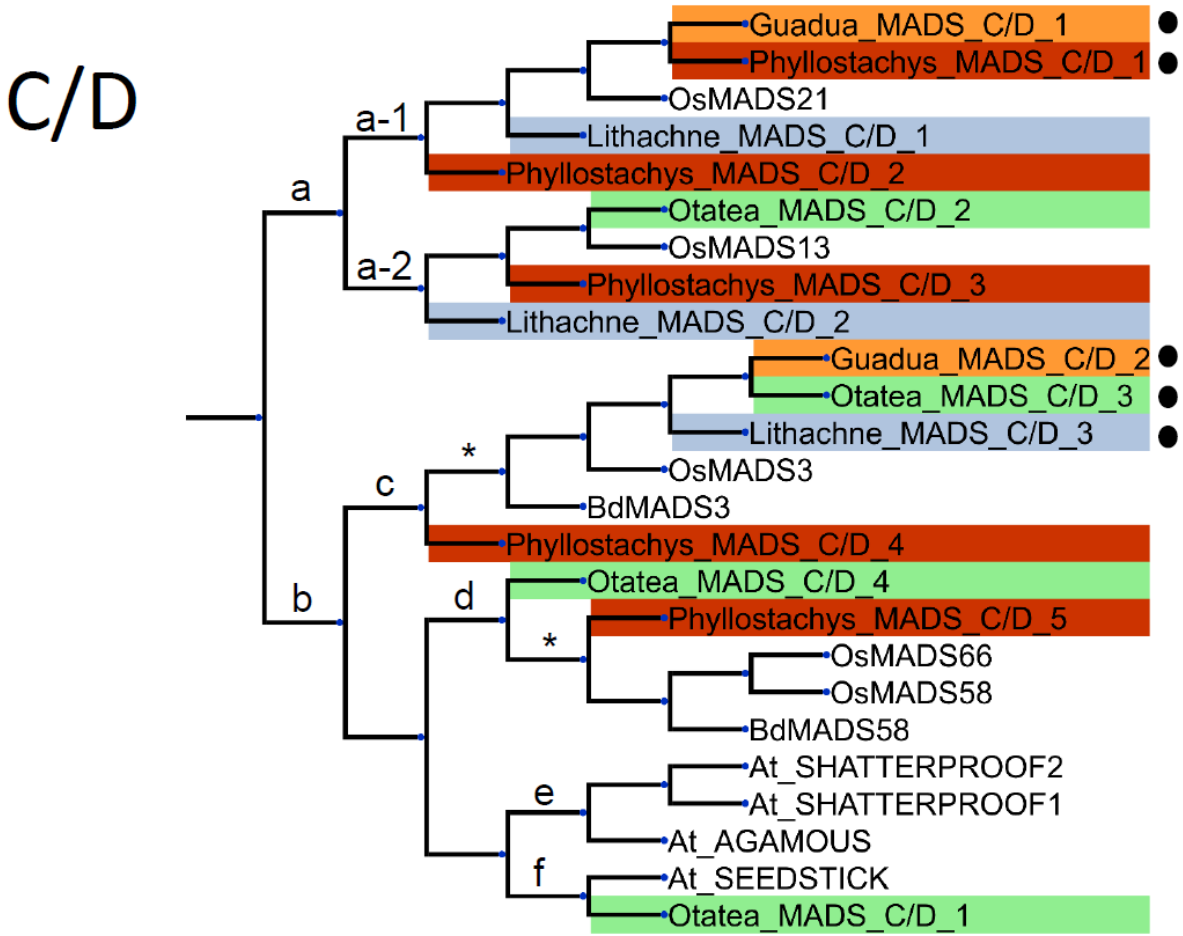
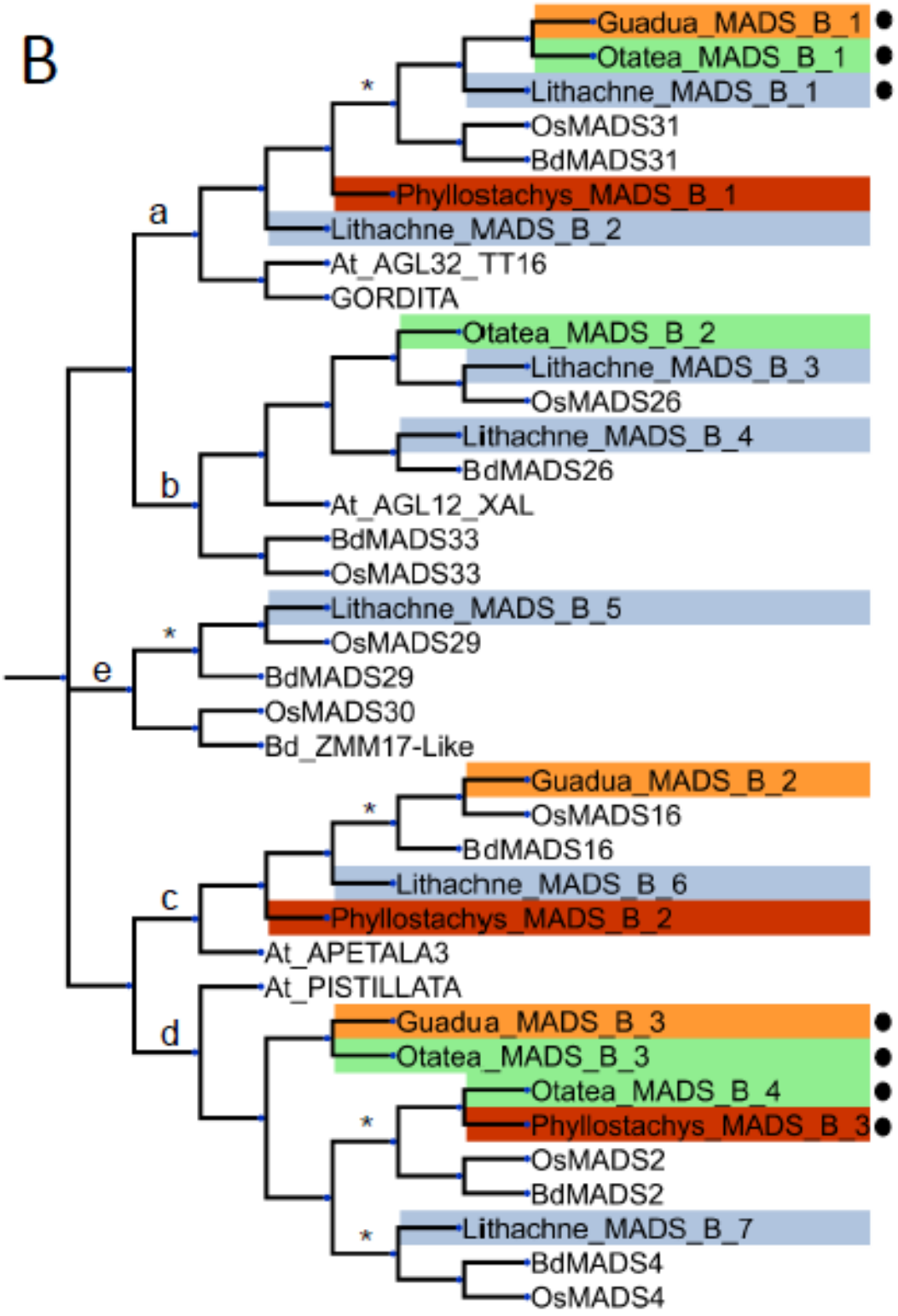


Figure 12. Neighbor-joining gene tree representing the A and C/D-class MADS box genes. Gene copies assembled in this study are labeled by genus, colored according to taxa (orange: *G. inermis*, green: *O. acuminata*, dark red: *P. aurea*, blue: *L. pauciflora*) and numbered redundantly to distinguish copies. Reference gene copies are not colored, are abbreviated by binomial (At: *Arabidopsis thaliana*, Bd: *Brachypodium distachyon*, Os: *Oryza sativa*) and are numbered according to their labeling in Genbank.


B
Guadua_MADS_B_1
Otatea_MADS_B_1
Lithachne_MADS_B_1
OsMADS31
BdMADS31
Phyllostachys_MADS_B_1
Lithachne_MADS_B_2
At_AGL32_TT16
GORDITA
Otatea_MADS_B_2
Lithachne_MADS_B_3
OsMADS26
Lithachne_MADS_B_4
BdMADS26
At_AGL12_XAL
BdMADS33
OsMADS33
Lithachne_MADS_B_5
OsMADS29
BdMADS29
OsMADS30
Bd_ZMM17-Like
Guadua_MADS_B_2
OsMADS16
BdMADS16
Lithachne_MADS_B_6
Phyllostachys_MADS_B_2
At_APETALA3
At_PISTILLATA
Guadua_MADS_B_3
Otatea_MADS_B_3
Otatea_MADS_B_4
Phyllostachys_MADS_B_3
OsMADS2
BdMADS2
Lithachne_MADS_B_7
BdMADS4
OsMADS4
a
b
c
d
e
*
*
*
*
*

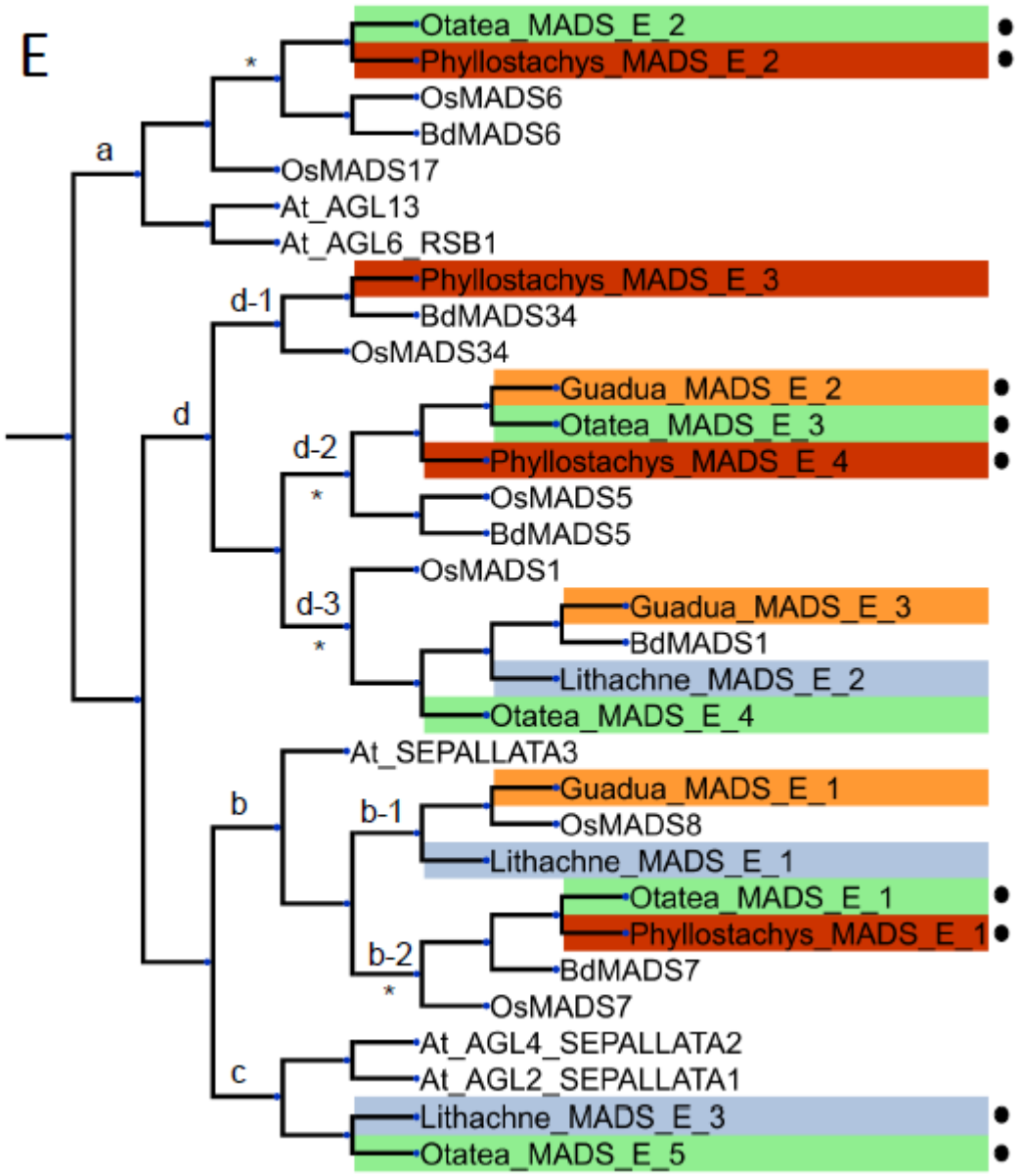
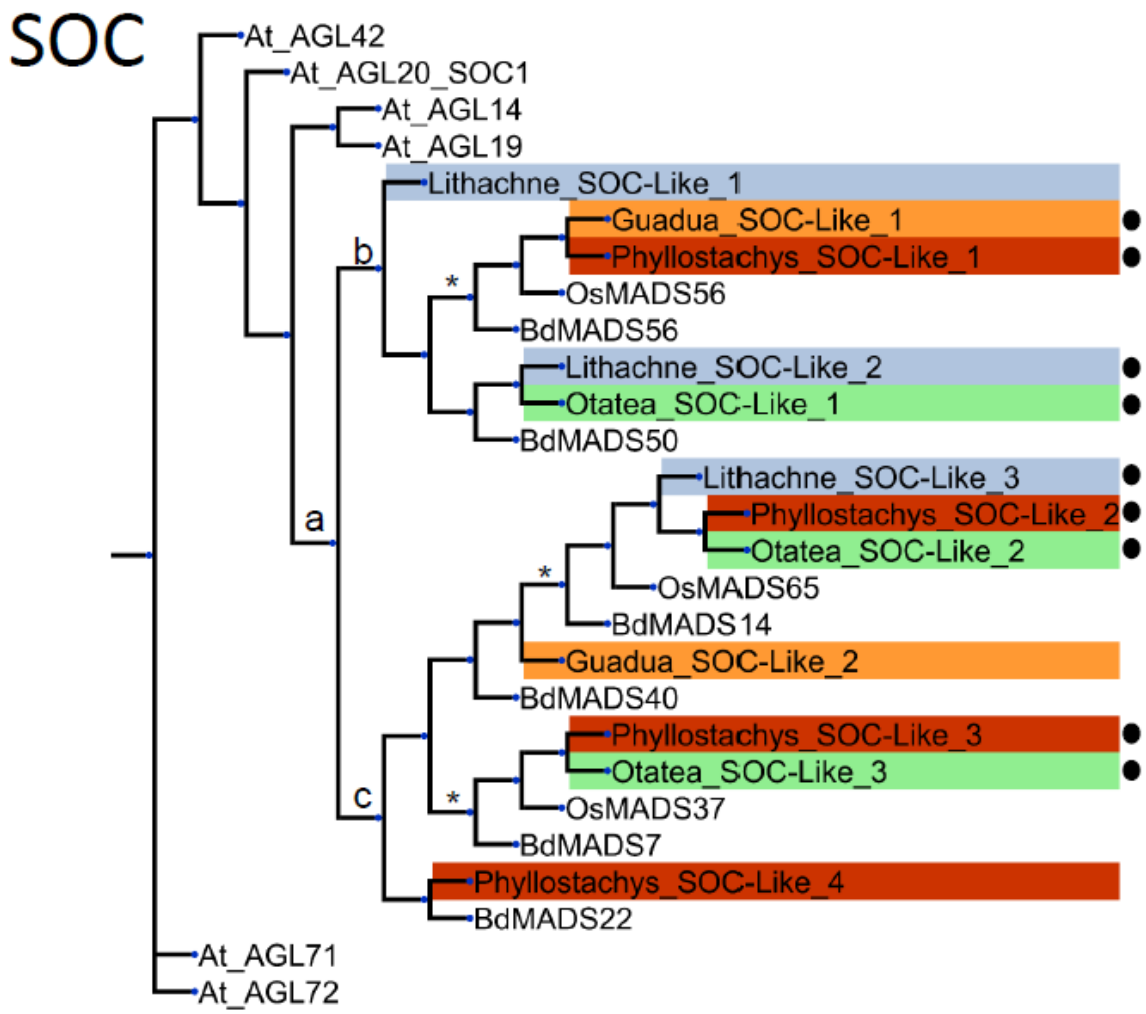


Figure 13. Neighbor-joining gene tree representing the B and E-class MADS box genes. Gene copies assembled in this study are labeled by genus, colored according to taxa (orange: *G. inermis*, green: *O. acuminata*, dark red: *P. aurea*, blue: *L. pauciflora*) and numbered redundantly to distinguish copies. Reference gene copies are not colored, are abbreviated by binomial (At: *Arabidopsis thaliana*, Bd: *Brachypodium distachyon*, Os: *Oryza sativa*) and are numbered according to their labeling in Genbank.


SOC
At_AGL42
At_AGL20_SOC1
At_AGL14
At_AGL19
Lithachne_SOC-Like_1
Guadua_SOC-Like_1
Phyllostachys_SOC-Like_1
b
*
OsMADS56
BdMADS56
Lithachne_SOC-Like_2
Otatea_SOC-Like_1
BdMADS50
Lithachne_SOC-Like_3
Phyllostachys_SOC-Like_2
Otatea_SOC-Like_2
a
*
OsMADS65
BdMADS14
Guadua_SOC-Like_2
BdMADS40
Phyllostachys_SOC-Like_3
Otatea_SOC-Like_3
c
*
OsMADS37
BdMADS7
Phyllostachys_SOC-Like_4
BdMADS22
At_AGL71
At_AGL72

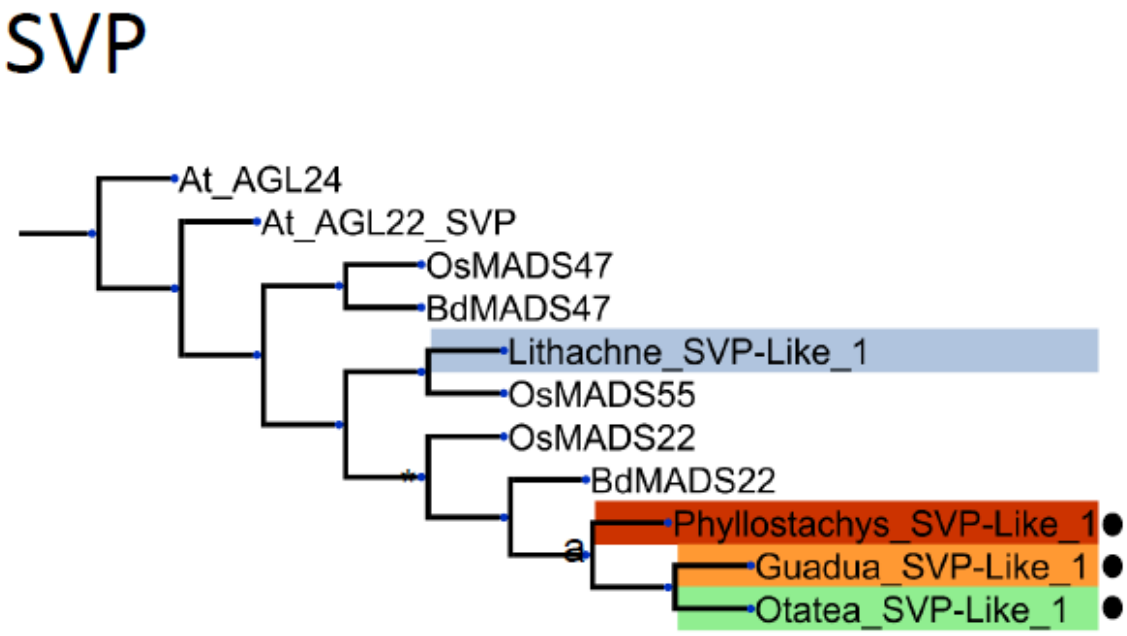


Figure 14. Neighbor-joining gene tree representing the *SOC* and *SVP*-like MADS box genes. Gene copies assembled in this study are labeled by genus, colored according to taxa (orange: *G. inermis*, green: *O. acuminata*, dark red: *P. aurea*, blue: *L. pauciflora*) and numbered redundantly to distinguish copies. Reference gene copies are not colored, are abbreviated by binomial (At: *Arabidopsis thaliana*, Bd: *Brachypodium distachyon*, Os: *Oryza sativa*) and are numbered according to their labeling in Genbank.

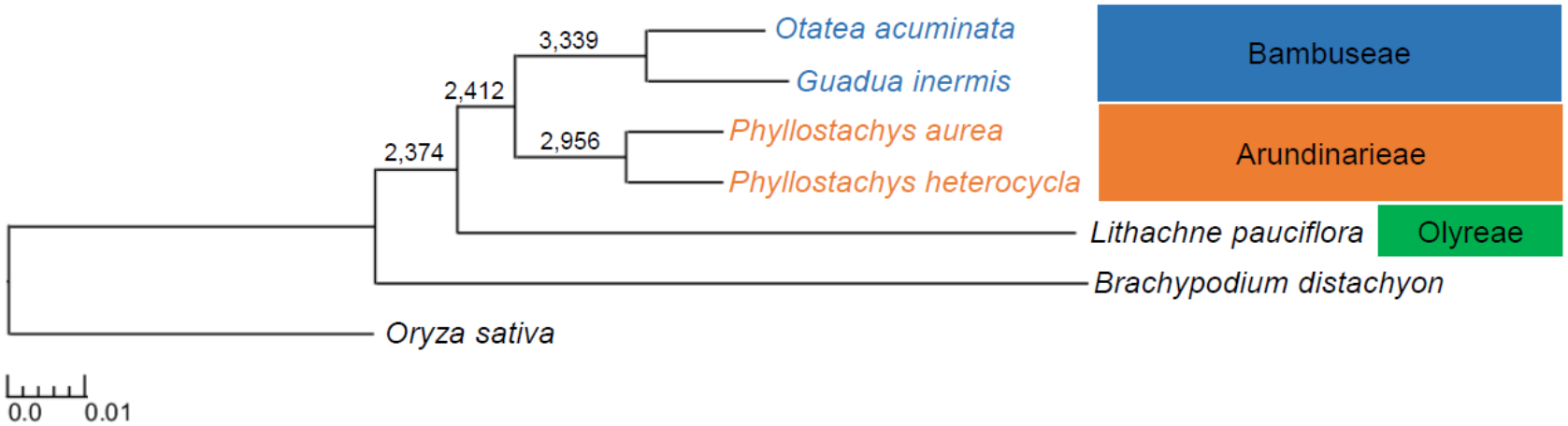


Figure 15. A maximum-likelihood tree generated using 3,878 concatenated gene alignments (2,698,410 bp). Branch lengths reflect number of substitution mutations. The tropical woody (Bambuseae), temperate woody (Arundinarieae) and herbaceous (Olyreae) bamboo tribes are indicated. Numbered nodes denote the number of the 3,878 best trees from each separate gene that support each node. All nodes were supported at 100% ML bootstrap.

CHAPTER 5

FIRST COMPLETE PLASTID GENOME FROM JOINVILLEACEAE (*J. ASCENDENS; POALES*) SHOWS UNIQUE AND UNPREDICTED REARRANGEMENTS

ABSTRACT

Joinvilleaceae is a family of tropical grass-like monocots that comprises only the genus *Joinvillea*. Previous studies have placed Joinvilleaceae in close phylogenetic proximity to the well-studied grass family. Here a full plastome sequence was determined and characterized for *J. ascendens.* The plastome was sequenced with next generation methods, fully assembled de novo and annotated. Plastid rearrangements were identified consistent with two novel inversions specific to the Joinvilleaceae lineage in addition to at least one novel plastid inversion in the Joinvilleaceae-Poaceae lineage. Two previously documented inversions in the Joinvilleaceae-Poaceae lineage and one previously documented inversion in the Poaceae lineage were also verified. Inversion events were verified visually and computationally. Additionally, the loss and subsequent degradation of the *accD* gene in Poales was explored extensively in Poaceae and *J. ascendens*.

INTRODUCTION

Among monocot Poales the grass family is most conspicuous ecologically and important agronomically. Close relatives to the grasses are found in what has been most recently

recognized as the graminid clade (Linder and Rudall 2005). The graminids also include the smaller families Joinvilleaceae, Ecdeiocoleaceae, Flagellariaceae, Restionaceae, Anarthriaceae and Centrolepidaceae, which are not as well studied, but suggest biogeographical, morphological, and phylogenomic aspects that bear on the evolutionary origins of the large grass radiation. Joinvilleaceae was selected as a representative "near-grass" for this study. The family is of particular interest here because of its previously reported and detailed plastome molecular evolutionary history (Doyle et al., 1992).

Joinvilleaceae includes the single genus *Joinvillea*, which includes four recognized species: *J. ascendens*, *J. borneensis*, *J. bryanii*, and *J. plicata*. These species are found on paleotropical islands of the Pacific (http://www.theplantlist.org/tpl/search?q=Joinvillea; http://www.catalogueoflife.org/; both accessed, 5 July 2015). *Joinvillea* species are robust grass-like herbs with strongly plicate leaves and flowers that are borne in large, multibranched terminal panicles. The small, perfect flowers themselves are distinguished from those of grasses in having a conspicuous perianth of six tepals and non-plumose stigmas, both of which suggest insect pollination, possibly by ants (Givnish et al., 2010). After fertilization the ovaries develop into fleshy drupes (Watson and Dalwitz, 1992 onwards), again contrasting with the caryopsis fruit type typical of Poaceae.

Most angiosperm plastid chromosomes (plastomes) are structured with two single copy regions, referred to as the large and small single copy (LSC and SSC) regions, flanked by two inverted perfectly repeated regions (IR) (Palmer and Stein, 1986). Compared to grasses, the plastomes of *Joinvillea* spp. were suggested to have undergone

major kilobase-sized rearrangements indicative of large scale inversion events (Doyle et al., 1992; Katayama and Ogihara, 1993; Michelangeli et al., 2003). PCR experiments primed from alternative sites in the large single copy region of the plastome led to these predictions. This result can now be re-explored with full plastome sequences obtained by next generation sequencing (NGS) technology.

While over 140 full grass plastomes have been assembled using traditional Sanger sequencing (Burke et al., 2012; Jones et al. 2014; Leseberg and Duvall, 2009; Morris and Duvall, 2010; and many others) and NGS (Burke et al., 2014; Cotton et al., 2015; Saarela et al., 2015; Wysocki et al., 2015), a complete plastome from other families within the graminid clade has yet to be sequenced. Here, the complete plastid chromosome of *Joinvillea ascendens* is sequenced and described. Evident in the arrangement of subgenomic sections, when compared to those of grasses and to the common cattail (*Typha latifolia*), are several large-scale molecular evolutionary events.

METHODS

DNA extraction and plastome sequencing

Silica-dried leaf material was provided by Dr. Lynn Clark, Iowa State University, Ames, IA under voucher (Clark & Attigala 1714). The tissue was homogenized in liquid nitrogen and DNA extraction was performed using the DNeasy plant minikit (Qiagen, Valencia, CA, USA) according to manufacturer instructions. Extractions were quantified using the Qubit fluorometric assay (Life Technologies, Grand Island, NY, USA) and diluted to 2.5 ng/ µL in 20 µL. Illumina libraries were prepared using the Nextera kit

(Illumina, San Diego, California, USA) and sequenced on a HiSeq 2000 platform at the Iowa State University DNA core facility (Ames, IA, USA).

Reads were assembled into contiguous sequences (contigs) using Velvet v. 1.2.08 (Zerbino and Birney, 2008) iteratively by using a range of kmer sizes (19--85 by steps of six) and running Velvet using assembled contigs as input (Wysocki et al., 2014). Contigs were then scaffolded by locating those that exhibit large regions of chloroplast homology and by extending them using *in silico* genome walking until perfect overlap of at least 20 base pairs (bp) with other contigs was obtained. This was repeated until the quadripartite plastome structure was completed. The plastome was annotated by aligning plastid genes from two reference plastomes: *Anomochloa marantoidea* (NC014062) and *Typha latifolia* (NC013823). *A. marantoidea* was used as the representative grass plastome because of its previous designation as a deeply diverging grass (GPWG, 2001). *T. latifolia* was used as a reference plastome for gene order (see below).

Aligning the plastome to other published monocot plastomes revealed a novel gene order. BLASTN (Altschul et al., 1997) was used to identify regions of significant sequence similarity between the *J. ascendens, A. marantoidea,* and *T. latifolia* plastomes. These regions were annotated and plastome rearrangements were first manually identified. The Mauve alignment algorithm (Darling et al., 2010) was then used to verify these rearrangements by identifying locally colinear blocks between these species in Geneious Pro version 7.1.8 (Biomatters Ltd., Auckland, New Zealand). All rearrangement boundaries were inspected manually for a continuum of overlap and read coverage to identify potential assembly errors.

Verification of rearrangement events

The observed rearrangement of plastome subregions was consistent with large scale inversions and events were hypothesized by visually inspecting BLASTN and Mauve output; however, a computational approach was executed to verify these inversions and to potentially uncover a more parsimonious series of events. A custom Python script, which was named Detection of Inversion Modes through the Simulation of Unifying Mutations (DIMSUM), was produced to perform a computational confirmation. The script randomly inverted subsets of defined regions (identified by Mauve) within a starting plastome 100 times until a predefined endpoint was reached. DIMSUM performed a user-defined number of replications of this calculation and reported all solutions that were reached. The Python script and documentation can be found at http://sourceforge .net/projects/grassplastome/.

RESULTS

Sequencing the Joinvillea ascendens plastome

After trimming and filtering, the Illumina library generated 8,982,874 single-end reads of 25--100 bp (mean length: 92.4 bp). The completed *J. ascendens* plastome was 149,327 bp in length. The short and large-single copy regions were 12,907 and 85,526 bp respectively with two inverted repeat regions of 25,447 bp each. The read depth of the plastome ranged from 19--147 with a mean depth of 77.1. The *J. ascendens* plastome had an AT composition of 60.4% and contained 117 protein coding sequences, 35 tRNA genes, and eight rRNA genes.

Identification of major inversions

Joinvillea ascendens and *A. marantoidea* shared a colinear plastome except for one large region in the LSC domain between *rps16* and *psaI*. This region was divided into two homologous subregions of sizes ~33 kbp (*rps14--trnQ-UUG*; Ja-LSC1) and ~19.5 kbp (psaB--rbcL; Ja-LSC2). The two had rearranged laterally with the Ja-LSC1 subregion retaining its orientation and the Ja-LSC2 subregion exhibiting reverse-complementation (Fig. 16).

A. marantoidea exhibited two large regions of rearrangement when compared to *T. latifolia*. One subregion within the LSC, between *trnfM--trnE* (~23 kbp; D-LSC1), was inverted in *A. marantoidea* and had laterally exchanged positions with the adjacent subregion, between *psbD--trnfM* (~5 kbp; D-LSC2). A third region that contained *trnG* in *T. latifolia* (~750 bp; D-LSC3) bordered the upstream boundary of the inversion in *T. latifolia* and was rearranged to a reverse-complemented position between the two LSC rearrangements (Fig. 17). The remaining portion of the LSC and IR regions exhibited colinearity. The SSC exhibited a similar pattern as the LSC with two subregions rearranged laterally with one region retaining its orientation (*ndhF--ndhH*, ~14 kbp; PoJo-SSC1) and the other reverse-complemented (*ndhH--ycf1*, ~3 kbp; PoJo-SSC2). This rearrangement was also present in all other grass subfamilies and in *J. ascendens*.

Unique plastome features

The *accD* gene, which encodes the beta-carboxyl transferase subunit of acetyl-CoA carboxylase (Kode et al., 2005), was not present as a functional copy in *J. ascendens*. In *J. ascendens*, *accD* was truncated to a pseudogene (ψ*accD*) approximately

66 aa upstream of the stop codon in *T. latifolia*. The upstream portion of ψ*accD* was also truncated by an inversion break-point. The *accD* gene was deleted from the *A. marantoidea* plastome completely.

The *ycf2* gene was also present in *T. latifolia* but absent from *A. marantoidea* except for two small pseudogene fragments of 360 and 507 bp. The *ycf2* gene was mostly absent from the *J. ascendens* plastome except for two pseudogene fragments of 368 and 1641 bp. The *ycf1* gene is found intact in *T. latifolia*, but in *J. ascendens* is found as a pseudogene that is truncated by approximately 1,399 aa due to an internal stop created by a nonsense mutation (UAU → UAG). Approximately half of the downstream region of *ycf1* has been deleted in *J. ascendens. A. marantoidea*, like *J. ascendens*, did not possess a functional copy of *ycf1*. The upstream portion of *ycf1* that was present in *J. ascendens* was deleted in *A. marantoidea.*

The *clpP* gene contained one intron in *J. ascendens* compared to two in *T. latifolia* and no introns in *A. marantoidea*. Additionally, a ~300 bp inversion between *A. marantoidea* and *T. latifolia* was observed in the *ycf3-trnS* intergenic spacer of the LSC.

DISCUSSION

Large plastome rearrangements

Previous studies that documented monocot-specific plastid rearrangements used the plastome from the eudicot *Nicotiana tabacum* to represent the ancestral state (Doyle et al. 1992; Michelangeli et al., 2003). Here, we used the monocot *T. latifolia* due to its relatively closer phylogenetic proximity to the taxa of interest and consequent higher

sequence similarity. Because our approach for inversion identification is based on homology inferred from sequence similarity, using *T. latifolia* as a reference point allows for a putatively more accurate inference of evolutionary events. A preliminary alignment of *T. latifolia* and *N. tabacum* showed no large scale differences in synteny between the two taxa. The phylogenetic distance between *T. latifolia* and *N. tabacum* also allows us to confidently infer that the gene order present in these two plastomes is ancestral.

The largest documented inversion in the plastid LSC from Doyle et al. (1992) that united Restionaceae, Joinvilleaceae, and Poaceae was verified in this study as the event that inverted the D-LSC1 and D-LSC2 regions of *J. ascendens* (Fig. 17; A). The second largest inversion, which united only Poaceae and Joinvilleaceae and which was documented in the same study, was also verified here as the event that returned D-LSC2 to the original orientation and inverted the adjacent D-LSC3 (Fig. 17; B). The third and comparably smaller inversion (< 300 bp; unnamed), which is specific to the grass family, was also verified here.

The *J. ascendens* plastome exhibits a pattern of rearrangement consistent with two large inversions in the LSC region. The first inversion occurred in Ja-LSC1 and Ja-LSC2 (Fig 16; C) and the second occurred in Ja-LSC2, which returned it to the original orientation (Fig 16; D). Because no evidence of these two inversions is found in grass plastomes, these rearrangements are likely to have occurred after the divergence of the Joinvilleaceae and the Ecdeiocoleaceae + Poaceae lineage (Fig. 18). Further sequencing of *Joinvillea* sp. plastomes would determine whether these rearrangements are

synapomorphic to all members of the family or if they are specific to *J. ascendens* or another subclade of *Joinvillea*.

The rearrangement of the SSC region in Joinvilleaceae and Poaceae may suggest the occurrence of two inversions events. The first event would have inverted both PoJo-SSC1 and PoJo-SSC2 with a subsequent inversion of PoJo-SSC1 to the original orientation. These two events would have had to occur before the divergence of the common ancestor of Joinvilleaceae-Poaceae. The phylogenetic hypothesis in which Joinvilleaceae has a sister relationship to Poaceae+Ecdeiocoleaceae (Givnish et al., 2010) would suggest that these inversion events should be evident in all members of Ecdeiocoleaceae. However, the exact placement of these inversions will remain obscured until sequences from additional monocot plastomes, such as *Flagellaria* sp., are determined. Alternatively, this pattern may be indicative of one inversion event. The actual orientation of any SSC region at a given time cannot be determined with certainty as it is flanked by two IR regions. Approaching the SSC from either orientation will produce identical sequencing results. Typically, SSC regions are assembled to be alignable to previously sequenced plastomes. The orientation of each single-copy region could be constantly in flux with homologous recombination of the IR regions (Palmer, 1983). Stein et al. (1986) suggested that single-copy regions exist in equimolar populations of the two orientations.

When compared to *T. latifolia*, the *J. ascendens* plastome seems to have accumulated a fairly confounding series of inversions. However, when compared to *A. marantoidea* it becomes clear that the LSC region rearrangements of *J. ascendens* are a

result of two large inversions that occurred in the Poaceae-Joinvilleaceae lineage followed by two large inversions that are specific to the Joinvilleaceae lineage. The two large inversions that were present in the Poaceae-Joinvilleaceae lineage are completely embedded in the ~33 kbp inversion specific to Joinvilleaceae, which explains why these inversions were not previously detected using PCR-based techniques with completely internal priming sites.

Verification of rearrangement events

All series of inversion were shown to be the most parsimonious by running 1,000,000 replicates of DIMSUM. Although these events were initially found through visual inspection, DIMSUM was used to easily elucidate inversion events between sequences with more confounding steps and fewer collinear regions. Even within sequences with easily defined events, DIMSUM located the shortest series of inversions among a large number of sequences quickly and identified large-scale evolutionary patterns.

Unique plastome features

No functional copies of *accD* were located in the remaining grass subfamilies, but remnants of the degraded copy were variably present. Anomochlooideae, and all subfamilies within the "PACMAD" (Panicoideae, Aristidoideae, Chloridoideae, Micrairoideae, Arundinoideae and Danthonioideae) clade did not contain any remnants of ψ*accD*. The remaining grass subfamilies did contain ψ*accD* sequences that ranged from 59--451 bp. The structure of the ψ*accD* exhibits differential degradation between grasses

and Joinvilleaceae as seen in pseudogenes or other noncoding regions in independent evolutionary lineages of grasses (Duvall et al. in press; Maier et al. 1995; Wysocki et al. 2015). An upstream region of ψ*accD* remained in *J. ascendens* and was deleted in all of the grasses and a downstream sequence remained in grasses but was deleted in *J. ascendens*. Remaining pseudogene sequences also suggested two losses of ψ*accD* in Poaceae with one loss in the Anomochlooideae and a second loss in the PACMAD clade. A nearly identical region of ψ*accD* (~450 bp) is present in Pharoideae and Puelioideae, which suggests that this pattern of degradation is plesiomorphic due to the position of these two subfamilies as a basal grade. One study (Harris et al., 2013) reported that a larger remnant of ψ*accD* was found in the Chloridoideae, specifically within the genus *Eragrostis*. Previously published full plastomes from *E. tef* (KT168385) and *E. minor* (KT168384) exhibited no such sequence.

The structure of *clpP* in these taxa shows variation in intron number across taxa. The *T. latifolia* plastome has two clpP introns, while *J. ascendens* and *A. marantoidea* have one and no introns respectively. This may be indicative of a stepwise loss of introns in the grasses with the first loss taking place in the Poaceae-Joinvilleaceae lineage followed by an independent loss in Poaceae. An alternative series of events, although less likely, could be the independent losses of one and two introns in Joinvilleaceae and Poaceae respectively.

Conclusions

Major structural rearrangements are relatively rare in the plastomes of angiosperms, and have been most thoroughly explored in the graminid Poales. A hypothesized cause for

large plastome inversions is recombination between dispersed tRNA loci with similar sequences (Hiratsuka et al., 1989). However, not all of the major inversions observed in graminid Poales are flanked by tRNA genes or pseudogenes, so other inversion mechanisms may also be involved. Instances of reversals to the original sequence orientation (e.g. Figs 16, 17) suggest that some type of persistent activating sequence feature may facilitate reversals. The role of selection has not been explored in these events, but preserving the proximity of promoters to their polycistronic plastid operons is a hypothetical selection pressure worth exploring. At present, the complete characterization of relatively rare structural rearrangements in the fully sequenced plastomes of graminid Poales serves to better define branch points in the phylogeny of the group and more precisely delimit evolutionary lineages.

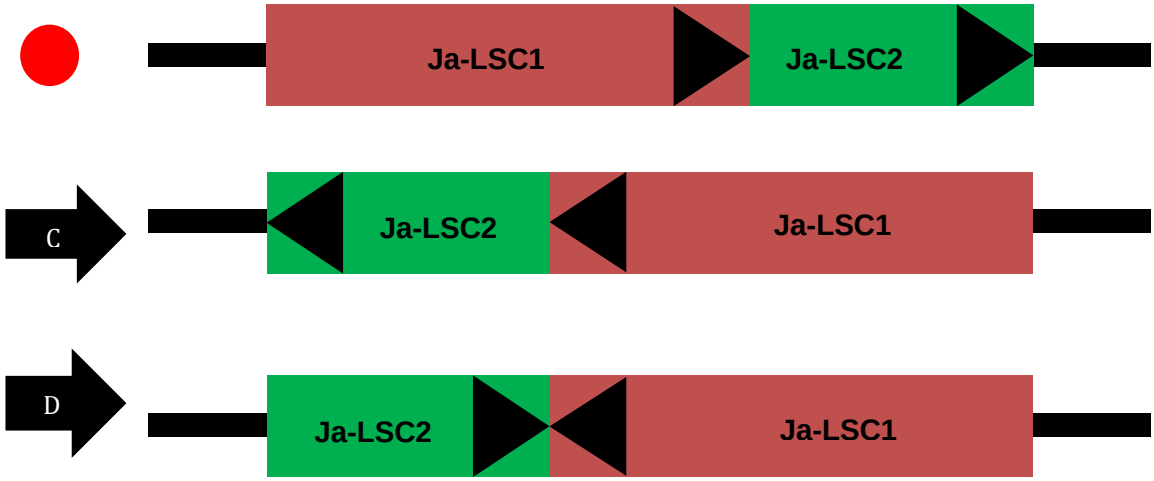


Figure 16. Diagram of the inversions that occurred in the *Joinvillea* lineage within the large single-copy subregion of the plastome. The red circle signifies the ancestral plastome before the divergence between Joinvilleaceae and Poaceae (see Fig. 18) and the arrows (C and D) represent large-scale inversion events. The bottom region represents the present arrangement of the *J. ascendens* plastome. Triangular markers are placed on each colored region to demonstrate orientation. Subregions are not drawn to scale.

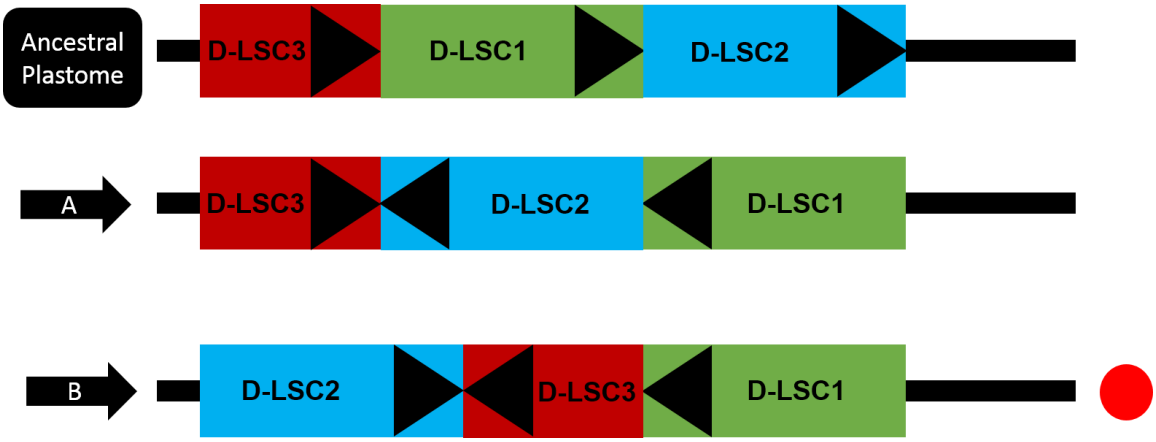


Figure 17. Diagram of the inversions that occurred in the Joinvilleaceae-Poaceae lineage within the large single-copy subregion of the plastome. The ancestral plastome signifies the pre-inversion state of the plastomes (*Typha latifolia*) and the red circle signifies the ancestral plastome before the divergence between Joinvilleaceae and Poaceae (see Fig. 18) and the arrows (A and B) represent large-scale inversion events. Triangular markers are placed on each colored region to demonstrate orientation. Subregions are not drawn to scale.

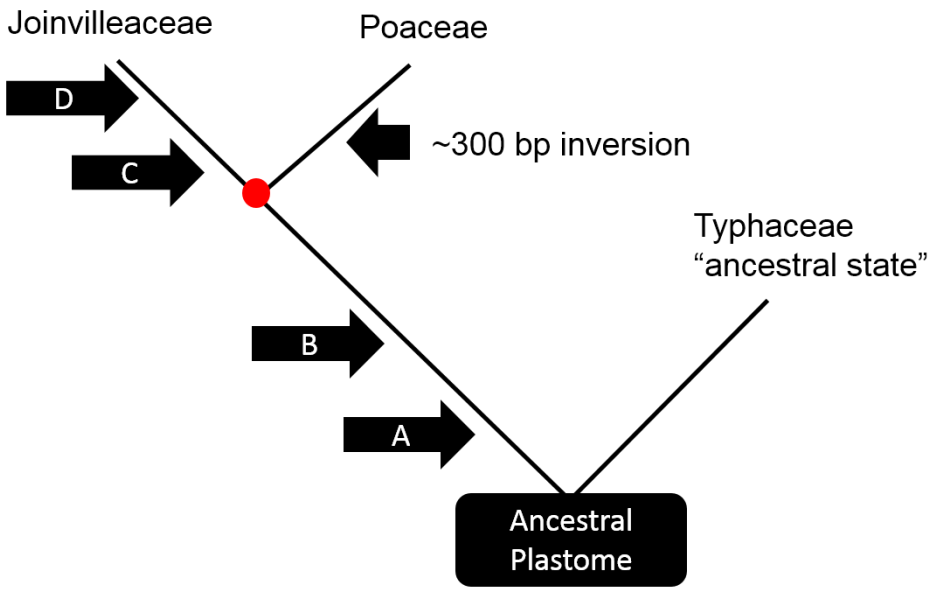


Figure 18. A simplified phylogenetic tree diagram with arrows that indicate the hypothesized relative position of each of the mutations from Fig. 16 and Fig. 17 (A--D) and one inversion exclusive to the grass lineage. Branch lengths are not to scale.

SYNTHESIS

Development of plastome assembly protocol

To date, the methods outlined in the first chapter of this dissertation are used regularly to successfully assemble grass plastomes. In this section I will outline the accepted protocol as of the date of the publication of this dissertation. The protocol starts with raw Illumina read files from the sequencing facility and ends with a complete and annotated plastome.

Raw reads are trimmed at each end according to their quality scores for each base pair and resulting reads less than 25 bp in length are removed. These steps allow for a cleaner continuum of overlap to be produced in the initial and downstream assembly steps. Next the reads are assembled using a de Bruijn-based de novo assembler such as Velvet or SPAdes. The resulting contigs are then scaffolded automatically using the ACRE method (see chapter 1). Overlapping reads and contigs are then used to manually extend each of the scaffolded contigs to fill gaps between them in a process called in silico genome walking. Genome walking can be automated and substantially accelerates the assembly process. However, attempts to automate this have not proven beneficial due to sequencing errors, the presence of untrimmed adapter sequences and sequencing of the correct inverted-repeat boundary. Human attention alleviates these problems and allows for sequencing/ assembly errors to be located at earlier steps in assembly.

When the full LSC and SSC regions and one IR have been assembled, BLASTN is used to locate inverted repeat boundaries and annotate the four subregions of the plastome (LSC, SSC, IRb, IRa) using a previously developed method (Burke et al. 2012). The raw reads are then mapped to the plastome for verification and theoretically should

produce a continuum of identical overlap. If this is not the case, the plastome is manually repaired according to the sequence given by the raw reads as they should be reflective of the actual sequence *in vivo*. The plastome is then aligned to a closely related plastome that had been previously annotated. Annotations are then transferred to the new plastome according to sequence similarity. After protein coding sequences are inspected for a valid start and stop codon, the new plastome is ready for downstream analysis and/or submission to Genbank.

Conclusions on bamboo phylogenetics

The plastome phylogenomic analyses performed in this dissertation strongly and confidently supported that woody bamboos are paraphyletic with herbaceous species. The temperate woody bamboos (Arundinarieae) exhibit a sister relationship to a clade including the herbaceous bamboos (Olyreae) and the tropical woody bamboos (Bambuseae) according to the phylogenetic signal produced by the maternally inherited plastome. The nuclear phylogenomic study performed in chapter 4 of this dissertation used nuclear transcripts from bamboos and produced a phylogeny in which woody bamboos are monophyletic. Neither of these incongruent results are necessarily erroneous, as different phylogenetic signals are produced by maternally and biparentally inherited markers. This difference in signal is consistent with a hybridization between progenitors of Bambuseae and Olyreae soon after the divergence of the three tribes. Hybridization between bamboo species has been observed in a number of taxa (McClure, 1966; Wong and Low, 2011) and have been observed to be consistent with other phylogenetic results (Triplett et al., 2014).

Plastome phylogenomics largely confirmed the subtribal classification within Olyreae (Oliveira et al., 2014; Soreng et al., 2015) and lineage classification within Arundinarieae (Attigala et al., 2014; Zeng et al., 2010). However, a few issues were raised within Bambuseae. Bambusinae was resolved as paraphyletic with several species exclusively sharing a clade with the Hickeliinae. *Neohouzeaua mekongensis*, a member of the Melocanninae, was placed deep within the Bambusinae. Because the plastome for this species was sequenced in duplicate, the well supported placement of *Neohouzeaua* is unambiguous. However, a propagated misidentification may be responsible for this relationship.

MADS-box genes and transcriptome sequencing in bamboos

In chapter 4 of this dissertation, a survey of floral timing and development genes and transcriptome assembly was performed. The de novo assembly produced fewer transcripts than the reference-based assembly, which was likely under-assembled. The MADS-box genes exhibited almost no connection to actual species phylogeny, but had some interesting insights into floral phenotypes among different species. Because this analysis was performed as a transcriptomic sequencing rather than a genomic sequencing, all MADS-box genes were likely not represented. Transcriptome sequencing will only produce expressed genes and will not sequence introns, but will produce a relative abundance of nuclear genes compared to the same level of sequencing within a full genomic extraction.

Large-scale inversions within the Poales

Sequencing and assembly of the full plastid sequence of *Joinvillea ascendens* allowed for two novel inversions to be identified that were not previously known. Three more inversions that were previously identified using PCR products were verified by comparing full plastomes from grasses and *J. ascendens* with the *Typha latifolia* plastome, which likely exhibited a gene order that was ancestral to the Poales. Changes in gene composition and structure regarding *accD* and *clpP* were also examined.

Future projects

A number of projects are logical continuations of this dissertation. The method employed in plastome phylogenetics worked well in bamboos and in resolving relationships within other grass subfamilies (Burke et al., in review; Cotton et al., 2015; Duvall et al., 2016; Saarela et al., 2015). Although these methods are optimized for grasses, small changes can be implemented to apply them to other angiosperm families. Preliminary work on *Helianthus* (Asteraceae) has yielded ten plastomes using these methods (Wysocki, unpublished data).

The entire plastome has been used in the bamboo plastome phylogenomics chapters of this dissertation, but more signal is required to resolve some of the relationships among the woody bamboos. Plastome indels and inversions can be used to add a relatively small number of informative sites to an analysis and do not require any additional sequencing. Additional sampling using RNA-Seq-derived transcripts would be a likely more robust method for resolving these relationships. Although this would be a costly project, NGS is becoming cheaper by the year. In the future, a full nuclear genome

sequencing project on ~100 species of bamboo would increase our understanding of bambusoid evolution immensely.

Determining the large-scale events that led to the plastome structure found in Poaceae and Joinvilleaceae could be applied to other angiosperm families. Once unifying plastome structures are identified, deep divergence patterns could be inferred through the analysis of inversion modes and the likely intermediate steps. Large-scale mutations could also be identified in nuclear chromosomes but would likely resolve smaller-scale taxonomic relationships.

LITERATURE CITED

Aird, D., Ross, M.G., Chen, W.S., Danielsson, M., Fennell, T., Russ, C., Jaffe, D.B., Nusbaum, C. and Gnirke, A., 2011. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome biol*, *12*(2), p.R18.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*, *25*(17), pp.3389-3402.

Attigala, L., Triplett, J.K., Kathriarachchi, H.S. and Clark, L.G., 2014. A new genus and a major temperate bamboo lineage of the Arundinarieae (Poaceae: Bambusoideae) from Sri Lanka based on a multi-locus plastid phylogeny. *Phytotaxa*, *174*(4), pp.187-205.

Attigala L, Wysocki WP, Triplett JK, Duvall MR, Clark LG: Phylogenetic reconstruction of Arundinarieae (Bambusoideae; Poaceae) based on plastome and low-copy nuclear gene analyses. Submitted to *Molecular Phylogenetics and Evolution.* February 2016.

Bamboo biodiversity [http://www.eeob.iastate.edu/research/bamboo/bamboo.html]

Bamboo Phylogeny Group. An updated tribal and subtribal classification for the Bambusoideae (Poaceae). In: Gielis J, Potters G, editors. Proc of the 9th World Bamboo Congress. Antwerp, Belgium: World Bamboo Organization; 2012. p. 3-27.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D. and Pyshkin, A.V., 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology*, *19*(5), pp.455-477.

Bateman A, Coin L, Durbin R, Finn RD, Hollich V, et al. 2004. The Pfam protein families database. Nucleic Acids Res, 32(suppl 1), D138-D141.

Bortiri, E., Coleman-Derr, D., Lazo, G.R., Anderson, O.D. and Gu, Y.Q., 2008. The complete chloroplast genome sequence of Brachypodium distachyon: sequence comparison and phylogenetic analysis of eight grass plastomes. *BMC Research Notes*, *1*(1), p.61.

Bouchenak-Khelladi, Y., Salamin, N., Savolainen, V., Forest, F., van der Bank, M., 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. *Molecular phylogenetics and evolution*, *47*(2), pp.488-505.

Bouchenak-Khelladi, Y., Verboom, G.A., Savolainen, V. and Hodkinson, T.R., 2010. Biogeography of the grasses (Poaceae): a phylogenetic approach to reveal

evolutionary history in geographical space and geological time. *Botanical Journal of the Linnean Society*, *162*(4), pp.543-557.

Burke, S.V., Clark, L.G., Triplett, J.K., Grennan, C.P. and Duvall, M.R., 2014. Biogeography and phylogenomics of new world Bambusoideae (Poaceae), revisited. *American journal of botany*, *101*(5), pp.886-891.

Burke, S.V., Grennan, C.P. and Duvall, M.R., 2012. Plastome sequences of two New World bamboos—Arundinaria gigantea and Cryptochloa strictiflora (Poaceae)—extend phylogenomic understanding of Bambusoideae. *American journal of botany*, *99*(12), pp.1951-1961.

Chokthaweepanich, H., 2014. Phylogenetics and evolution of the paleotropical woody bamboos (Poaceae: Bambusoideae: Bambuseae).

Clark, L.G. and Triplett, J.K., 1993. Arundinaria. *Flora of North America Editorial Committee (eds.)*, pp.17-20.

Clark, L.G., 1989. Systematics of Chusquea Section Swallenochloa, Section Verticillatae, Section Serpentes, and Section Longifoliae (Poaceae-Bambusoideae). *Systematic Botany Monographs*, pp.1-127.

Clark LG, Zhang W, Wendel JF. 1995. A phylogeny of the grass family (Poaceae) based on ndhF sequence data. *Systematic Botany*, 436-460.

Clark LG, Londoño X, Ruiz-Sanchez E. 2015. Bamboo taxonomy and habitat. In: Laslo, P., Köehl, M (eds). Bamboo. Series: Tropical Forestry Handbook. Springer (In press).

Coen ES, Meyerowitz EM. 1991. The war of the whorls: genetic interactions controlling flower development. Nature, 353(6339): 31-37.

Coleman, P.J., 1980. Plate tectonics background to biogeographic development in the southwest Pacific over the last 100 million years. *Palaeogeography, Palaeoclimatology, Palaeoecology*, *31*, pp.105-121.

Cosner, M.E., Raubeson, L.A. and Jansen, R.K., 2004. Chloroplast DNA rearrangements in Campanulaceae: phylogenetic utility of highly rearranged genomes. *BMC Evolutionary Biology*, *4*(1), p.27.

Cotton, J. L., Wysocki, W. P., Clark, L. G., Kelchner, S. A., Pires, J. C., Edger, P. P., ... & Duvall, M. R. (2015). Resolving deep relationships of PACMAD grasses: a phylogenomic approach. BMC plant biology, 15(1), 178.

Cox, M.P., Peterson, D.A. and Biggs, P.J., 2010. SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. *BMC bioinformatics*, *11*(1), p.485.

Cronn, R., Liston, A., Parks, M., Gernandt, D.S., Shen, R. and Mockler, T., 2008. Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. *Nucleic acids research*, *36*(19), pp.e122-e122.

Darling, A. E., Mau, B., & Perna, N. T. (2010). progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PloS one, 5(6), e11147.

Darriba, D., Taboada, G.L., Doallo, R. and Posada, D., 2012. jModelTest 2: more models, new heuristics and parallel computing. *Nature methods*, *9*(8), pp.772-772.

Dhingra, A. and Folta, K.M., 2005. ASAP: amplification, sequencing & annotation of plastomes. *BMC genomics*, *6*(1), p.176.

Doyle, J.J., Davis, J.I., Soreng, R.J., Garvin, D. and Anderson, M.J., 1992. Chloroplast DNA inversions and the origin of the grass family (Poaceae). *Proceedings of the National Academy of Sciences*, *89*(16), pp.7722-7726.

Dransfield, S., 2002. Greslania circinata and Greslania rivularis (Poaceae-Bambusoideae) from New Caledonia. *BAMBOO SCIENCE & CULTURE*, p.1.

Dreni L, Jacchia S, Fornara F, Fornari M, Ouwerkerk PB et al. 2007. The D-lineage MADS-box gene OsMADS13 controls ovule identity in rice. *The Plant Journal*, 52(4): 690-699.

Duvall, M.R., Leseberg, C.H., Grennan, C.P. and Morris, L.M., 2010. Molecular evolution and phylogenetics of complete chloroplast genomes in Poaceae. *Diversity, phylogeny, and evolution in the monocotyledons. Aarhus University Press, Aarhus*, pp.437-450.

Felsenstein, J., 2005. PHYLIP Seattle: Department of Genome Science. *University of Washington*, *3*.

Finn RD, Clements J, Eddy SR. 2011. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res, gkr367.

Gao J, Zhang Y, Zhang C, Qi F, Li X, et al. 2014. Characterization of the Floral Transcriptome of Moso Bamboo (*Phyllostachys edulis*) at Different Flowering Developmental Stages by Transcriptome Sequencing and RNA-Seq Analysis. PloS One, 9(6), e98910.

Gaut, B.S., Clark, L.G., Wendel, J.F. and Muse, S.V., 1997. Comparisons of the molecular evolutionary process at rbcL and ndhF in the grass family (Poaceae). *Molecular Biology and Evolution*, *14*(7), pp.769-777.

Givnish, T. J., Ames, M., McNeal, J. R., McKain, M. R., Steele, P. R., dePamphilis, C. W.,Leebens-Mack, J. H. (2010). Assembling the tree of the monocotyledons: plastome sequence phylogeny and evolution of Poales 1. Annals of the Missouri Botanical Garden, 97(4), 584-616.

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 systematics and evolution*, *299*(1), pp.239-257.

Goldman, N., Anderson, J.P. and Rodrigo, A.G., 2000. Likelihood-based tests of topologies in phylogenetics. *Systematic biology*, *49*(4), pp.652-670.

Goremykin, V.V., Salamini, F., Velasco, R. and Viola, R., 2009. Mitochondrial DNA of Vitis vinifera and the issue of rampant horizontal gene transfer. *Molecular biology and evolution*, *26*(1), pp.99-110.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, et al. 2011. Full-length transcriptome assembly from RNA-seq data without a reference genome. Nat Biotech. 29(7): 644-52. doi: 10.1038/nbt.1883.

Grandcolas, P., Murienne, J., Robillard, T., Desutter-Grandcolas, L., Jourdan, H., Guilbert, E. and Deharveng, L., 2008. New Caledonia: a very old Darwinian island?. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1508), pp.3309-3317.

Grass Phylogeny Working Group [GPWG]. 2001. Phylogeny and subfamilial classificationof the grasses (Poaceae). Ann Missouri Bot Gard 88:373–45

Guerreiro, C. and Agrasar, Z. 2013. Two new species of Chusquea (Poaceae, Bambuseae) from northwestern Argentina. *Systematic Botany*, *38*(2), pp.390-397.

Guindon, S. and Gascuel, O., 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic biology*, *52*(5), pp.696-704.

Hand, M.L., Spangenberg, G.C., Forster, J.W. and Cogan, N.O., 2013. Plastome sequence determination and comparative analysis for members of the *Lolium-Festuca* grass species complex. *G3: Genes| Genomes| Genetics*, *3*(4), pp.607-616.

Harris, M. E., Meyer, G., Vandergon, T., & Vandergon, V. O. (2013). Loss of the acetyl-CoA carboxylase (*accD*) gene in Poales. Plant Molecular Biology Reporter, 31(1), 21-31.

Hartmann U, Höhmann S, Nettesheim K, Wisman E, Saedler H, Huijser P. 2000. Molecular cloning of *SVP*: a negative regulator of the floral transition in Arabidopsis. *The Plant Journal*, 21(4): 351-360.

Hernandez, D., François, P., Farinelli, L., Østerås, M. and Schrenzel, J., 2008. De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer. *Genome research*, *18*(5), pp.802-809.

Hisamoto, Y., Kashiwagi, H. and Kobayashi, M., 2008. Use of flowering gene *FLOWERING LOCUS T* (*FT*) homologs in the phylogenetic analysis of bambusoid and early diverging grasses. *Journal of plant research*, *121*(5), pp.451-461.

Huerta-Cepas J, Dopazo J, Gabaldón T. 2010. ETE: a python Environment for Tree Exploration. *BMC bioinformatics*, 11(1), 24.

Hunziker JH, Wulff AF, Soderstrom TR. 1982. Chromosome studies on the Bambusoideae (Gramineae). Brittonia,

Huson, D.H. and Bryant, D., 2006. Application of phylogenetic networks in evolutionary studies. *Molecular biology and evolution*, *23*(2), pp.254-267.

Janzen, D.H., 1976. Why bamboos wait so long to flower. *Annual Review of Ecology and Systematics*, pp.347-391.

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 systematics and evolution*, *300*(6), pp.1421-1436.

Judziewicz, E.J., Clark, L.G., Londono, X. and Stern, M.J., 1999. American bamboos. *American bamboos*.

Katayama, H., & Ogihara, Y. (1993). Structural alterations of the chloroplast genome found in grasses are not common in monocots. Current genetics, 23(2), 160-165.

Katoh, K., Kuma, K.I., Toh, H. and Miyata, T., 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic acids research*, *33*(2), pp.511-518.

Kelchner, S.A. and Group, B.P., 2013. Higher level phylogenetic relationships within the bamboos (Poaceae: Bambusoideae) based on five plastid markers. *Molecular phylogenetics and evolution*, *67*(2), pp.404-413.

Kellogg, EA. 2015a. IV. Subfamily Ehrhartoideae Link (1827). In Flowering Plants. Monocots (pp. 143-150). Springer International Publishing.

Kellogg, EA. 2015b. V. Subfamily Bambusoideae Luerss.(1893). In Flowering Plants. Monocots (pp. 151-198). Springer International Publishing.

Kellogg, EA. 2015c. VI. Subfamily Pooideae Benth.(1861). In Flowering Plants. Monocots (pp. 199-265). Springer International Publishing.

Kode, V., Mudd, E. A., Iamtham, S., & Day, A. (2005). The tobacco plastid accD gene is essential and is required for leaf development. The plant journal,44(2), 237-244.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nature Methods, 9(4), 357-359.

Lee J, Lee I. 2010. Regulation and function of SOC1, a flowering pathway integrator. *Journal of Experimental Botany*, 61(9): 2247-2254.

Lee JH, Park SH, Ahn JH. 2012. Functional conservation and diversification between rice OsMADS22/OsMADS55 and *Arabidopsis SVP* proteins. *Plant science*, 185: 97-104.

Leseberg, C.H. and Duvall, M.R., 2009. The complete chloroplast genome of *Coix lacryma-jobi* and a comparative molecular evolutionary analysis of plastomes in cereals. *Journal of Molecular Evolution*, *69*(4), pp.311-318.

Levy AA, Feldman M. 2002. The impact of polyploidy on grass genome evolution. Physiol, 130(4):

Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., Li, Y., Li, S., Shan, G., Kristiansen, K. and Li, S., 2010. De novo assembly of human genomes with massively parallel short read sequencing. *Genome research*, *20*(2), pp.265-272.

Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics, 22(13), 1658-1659.

Linder, H.P. and Rudall, P.J., 2005. Evolutionary history of Poales. Annual Review of Ecology, Evolution, and Systematics, pp.107-124.

Lima, R.A., Rother, D.C., Muler, A.E., Lepsch, I.F. and Rodrigues, R.R., 2012. Bamboo overabundance alters forest structure and dynamics in the Atlantic Forest hotspot. *Biological Conservation*, *147*(1), pp.32-39.

Lin CS, Lin CC, Chang WC. 2005. Shoot regeneration, re-flowering and post flowering survival in bamboo inflorescence culture. Plant Cell, Tissue and Organ Culture, 82(3), 243-249.

Lin, Y., Li, J., Shen, H., Zhang, L. and Papasian, C.J., 2011. Comparative studies of de novo assembly tools for next-generation sequencing technologies. *Bioinformatics*, *27*(15), pp.2031-2037.

Liu M, Qiao G, Jiang J, Yang H, Xie L, et al. 2012. Transcriptome sequencing and de novo analysis for ma bamboo (*Dendrocalamus latiflorus* Munro) using the Illumina platform. PloS One, 7(10), e46766.

Liu, X., Pande, P.R., Meyerhenke, H. and Bader, D.A., 2013. PASQUAL: Parallel techniques for next generation genome sequence assembly. *Parallel and Distributed Systems, IEEE Transactions on*, *24*(5), pp.977-986.

Lovett, J.C., 1994. Notes on moist forest bamboos and bambusoid grasses in eastern Tanzania. *East Afr Nat Hist Soc Bull*, *24*(1), pp.2-5.

Lowry, P.P., 1998. Diversity, endemism, and extinction in the flora of New Caledonia: a review. In *Proc Int Symp on Rare, Threatened, and Endangered Floras of Asia and the Pacific. Monograph ed. Taipei, Taiwan: Institute of Botany, Academica Sinica* (pp. 181-206).

Ma, P.F., Guo, Z.H. and Li, D.Z., 2012. Rapid sequencing of the bamboo mitochondrial genome using Illumina technology and parallel episodic evolution of organelle genomes in grasses. *PLoS One*, *7*(1), p.e30297.

Ma, P.F., Zhang, Y.X., Zeng, C.X., Guo, Z.H. and Li, D.Z., 2014. Chloroplast phylogenomic analyses resolve deep-level relationships of an intractable bamboo tribe Arundinarieae (Poaceae). *Systematic biology*, p.syu054.

Ma, P.F., Zhang, Y.X., Guo, Z.H. and Li, D.Z., 2015. Evidence for horizontal transfer of mitochondrial DNA to the plastid genome in a bamboo genus. *Scientific reports*, *5*.

Marchesini, V.A., Sala, O.E. and Austin, A.T., 2009. Ecological consequences of a massive flowering event of bamboo (Chusquea culeou) in a temperate forest of Patagonia, Argentina. *Journal of Vegetation Science*, *20*(3), pp.424-432.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. journal, 17(1), pp-10.

McClure, F.A., 1966. The bamboos. A fresh perspective. *The bamboos. A fresh perspective.*

Michelangeli, F. A., Davis, J. I., & Stevenson, D. W. (2003). Phylogenetic relationships among Poaceae and related families as inferred from morphology, inversions in the plastid genome, and sequence data from the mitochondrial and plastid genomes. American Journal of Botany, 90(1), 93-106.

Miller, J.R., Koren, S. and Sutton, G., 2010. Assembly algorithms for next-generation sequencing data. *Genomics*, *95*(6), pp.315-327.

Morris, L.M. and Duvall, M.R., 2010. The chloroplast genome of Anomochloa marantoidea (Anomochlooideae; Poaceae) comprises a mixture of grass-like and unique features. *American Journal of Botany*, *97*(4), pp.620-627.

Nowrousian, M., Stajich, J.E., Chu, M., Engh, I., Espagne, E., Halliday, K., Kamerewerd, J., Kempken, F., Knab, B., Kuo, H.C. and Osiewacz, H.D., 2010. De novo assembly of a 40 Mb eukaryotic genome from short sequence reads: *Sordaria macrospora*, a model organism for fungal morphogenesis. *PLoS Genet*, *6*(4), p.e1000891.

Oliveira, R.P., Clark, L.G., Schnadelbach, A.S., Monteiro, S.H., Borba, E.L., Longhi-Wagner, H.M. and van den Berg, C., 2014. A molecular phylogeny of *Raddia* and its allies within the tribe Olyreae (Poaceae, Bambusoideae) based on noncoding plastid and nuclear spacers. *Molecular phylogenetics and evolution*, *78*, pp.105-117.

Pabón-Mora, N., Wong, G. K. S., & Ambrose, B. A. (2014). Evolution of fruit development genes in flowering plants. Frontiers in plant science, 5.

Palmer, J.D., 1983. Chloroplast DNA exists in two orientations. Nature.

Palmer, J.D. and Stein, D.B., 1986. Conservation of chloroplast genome structure among vascular plants. Current genetics, 10(11), pp.823-833.

Parks, M., Cronn, R. and Liston, A., 2009. Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes. *BMC biology*, *7*(1), p.84.

Parks, M., Cronn, R. and Liston, A., 2012. Separating the wheat from the chaff: mitigating the effects of noise in a plastome phylogenomic data set from Pinus L.(Pinaceae). *BMC Evolutionary Biology*, *12*(1), p.100.

Paszkiewicz, K. and Studholme, D.J., 2010. De novo assembly of short sequence reads. *Briefings in bioinformatics*, p.bbq020.

Pelaz S, Ditta GS, Baumann E, Wisman E, Yanofsky MF 2000. B and C floral organ identity functions require *SEPALLATA* MADS-box genes. Nature, 405(6783): 200-203.

Peng Z, Zhang C, Zhang Y, Hu T, Mu S, et al. 2013. Transcriptome sequencing and analysis of the fast growing shoots of Moso bamboo (*Phyllostachys edulis*). PloS One, 8(11), e78944.

Peng Z, Lu Y, Li L, Zhao Q, Feng Q, et al. 2013. The draft genome of the fast-growing non-timber forest species moso bamboo (*P. heterocycla*). Nature Genet, 45(4), 456-461.

Renzaglia, K.S., Schuette, S., Duff, R.J., Ligrone, R., Shaw, A.J., Mishler, B.D. and Duckett, J.G., 2007. Bryophyte phylogeny: advancing the molecular and morphological frontiers. *The Bryologist*, *110*(2), pp.179-213.

Rice, D.W. and Palmer, J.D., 2006. An exceptional horizontal gene transfer in plastids: gene replacement by a distant bacterial paralog and evidence that haptophyte and cryptophyte plastids are sisters. *BMC biology*, *4*(1), p.31.

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D.L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M.A. and Huelsenbeck, J.P., 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic biology*, *61*(3), pp.539-542.

Ruiz-Sanchez, E., 2011. Biogeography and divergence time estimates of woody bamboos: insights in the evolution of Neotropical bamboos. *Boletín de la Sociedad Botánica de México*, *88*, pp.67-75.

Ryu CH, Lee S, Cho LH, Kim SL, Lee YS, et al. 2009. *OsMADS50* and *OsMADS56* function antagonistically in regulating long day (LD)-dependent flowering in rice. *Plant, cell & environment*, 32(10): 1412-1427.

Saarela, J. M., Wysocki, W. P., Barrett, C. F., Soreng, R. J., Davis, J. I., Clark, L. G., ... & Duvall, M.R. (2015). Plastid phylogenomics of the cool-season grass subfamily: Clarification of relationships among early-diverging tribes. AoB plants, plv046.

Sarma, V.V., 2009. Flowering of Melocanna baccifera (Bambusaceae) in northeastern India. *Curr Sci*, *96*(9), p.1165.

Schnable JC, Freeling M, Lyons E. 2012. Genome-wide analysis of syntenic gene deletion in the grasses. Genome Biol Evol, 4(3): 265-277.

Shikanai, T., Shimizu, K., Ueda, K., Nishimura, Y., Kuroiwa, T., & Hashimoto, T. (2001). The chloroplast clpP gene, encoding a proteolytic subunit of ATP-dependent protease, is indispensable for chloroplast development in tobacco.Plant and Cell Physiology, 42(3), 264-273.

Shimodaira, H. and Hasegawa, M., 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Molecular biology and evolution*, *16*, pp.1114-1116.

Skinner DJ, Hill TA, Gasser CS. 2004. Regulation of ovule development. *The Plant Cell*, 16: S32-S45.

Skøt, L., Sanderson, R., Thomas, A., Skøt, K., Thorogood, D., Latypova, G., Asp, T. and Armstead, I., 2011. Allelic variation in the perennial ryegrass FLOWERING LOCUS T gene is associated with changes in flowering time across a range of populations. *Plant Physiology*, *155*(2), pp.1013-1022.

Soderstrom, T.R. and Zuloaga, F.O., 1989. A revision of the genus Olyra and the new segregate genus *Parodiolyra* (Poaceae: Bambusoideae: Olyreae). *Smithsonian Contrib. Bot*, (69), pp.1-79.

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). *Journal of Systematics and Evolution*, *53*(2), pp.117-137.

Stamatakis, A., 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, *22*(21), pp.2688-2690.

Steele, P.R., Hertweck, K.L., Mayfield, D., McKain, M.R., Leebens-Mack, J. and Pires, J.C., 2012. Quality and quantity of data recovered from massively parallel sequencing: examples in Asparagales and Poaceae. *American Journal of Botany*, *99*(2), pp.330-348.

Stein, D.B., Palmer, J.D. and Thompson, W.F., 1986. Structural evolution and flip-flop recombination of chloroplast DNA in the fern genus *Osmunda*. Current genetics, 10(11), pp.835-841.

Straub, S.C., Cronn, R.C., Edwards, C., Fishbein, M. and Liston, A., 2013. Horizontal transfer of DNA from the mitochondrial to the plastid genome and its subsequent evolution in milkweeds (Apocynaceae). *Genome biology and evolution*, *5*(10), pp.1872-1885.

Sungkaew, S., Stapleton, C.M., Salamin, N. and Hodkinson, T.R., 2009. Non-monophyly of the woody bamboos (Bambuseae; Poaceae): a multi-gene region phylogenetic analysis of Bambusoideae ss. *Journal of plant research*, *122*(1), pp.95-108.

Swofford, D.L., 2003. {PAUP*. Phylogenetic analysis using parsimony (* and other methods). Version 4.}.

Theißen G, Saedler H. 2001. Plant biology: floral quartets. *Nature,* 409(6819): 469-471.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res, 22(22): 4673-4680.

Trapnell C, Pachter L, Salzberg SL. 2009. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics, 25(9), 1105-1111.

Triplett, J.K. and Clark, L.G., 2010. Phylogeny of the temperate bamboos (Poaceae: Bambusoideae: Bambuseae) with an emphasis on Arundinaria and allies. *Systematic Botany*, *35*(1), pp.102-120.

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 Phytologist*, *204*(1), pp.66-73.

Wang L, Park HJ, Dasari S, Wang S, Kocher JP et al. 2013. CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. Nucleic Acids Res, 41(6), e74-e74.

Wang Z, Gerstein M, & Snyder M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nature Rev Genet, 10(1), 57-63.

Watson, L., Dallwitz, M.J. and Johnston, C.R., 1986. Grass genera of the world: 728 detailed descriptions from an automated database. Australian Journal of Botany, 34(2), pp.223-230.

Wei B, Zhang RZ, Guo JJ, Liu DM, Li AL, et al. 2014. Genome-wide analysis of the MADS-box gene family in Brachypodium distachyon. PloS one, 9(1).

Whipple CJ, Ciceri P, Padilla CM, Ambrose BA, Bandong SL, Schmidt RJ. 2004. Conservation of B-class floral homeotic gene function between maize and Arabidopsis. *Development,* 131(24), 6083-6091.

Whittall, J.B., Syring, J., Parks, M., Buenrostro, J., Dick, C., Liston, A. and Cronn, R., 2010. Finding a (pine) needle in a haystack: chloroplast genome sequence divergence in rare and widespread pines. *Molecular Ecology*, *19*(s1), pp.100-114.

Wong, K.M. and Low, Y.W., 2011. Hybrid zone characteristics of the intergeneric hybrid bamboo x Gigantocalamus maplenensis (Poaceae: Bambusoideae) in peninsular Malaysia. *Garden Bull Singapore*, *63*, pp.375-83.

Wu, F.H., Kan, D.P., Lee, S.B., Daniell, H., Lee, Y.W., Lin, C.C., Lin, N.S. and Lin, C.S., 2009. Complete nucleotide sequence of Dendrocalamus latiflorus and Bambusa oldhamii chloroplast genomes. *Tree physiology*, p.tpp015.

Wu, Z.Q. and Ge, S., 2012. The phylogeny of the BEP clade in grasses revisited: evidence from the whole-genome sequences of chloroplasts. *Molecular Phylogenetics and Evolution*, *62*(1), pp.573-578.

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 evolutionary biology*, *15*(1), p.50.

Wysocki, W.P., Clark, L.G., Kelchner, S.A., Burke, S.V., Pires, J.C., Edger, P.P., Mayfield, D.R., Triplett, J.K., Columbus, J.T., Ingram, A.L. and Duvall, M.R., 2014. A multi-step comparison of short-read full plastome sequence assembly methods in grasses. *Taxon*, *63*(4), pp.899-910.

Wysocki WP, Ruiz-Sanchez E, Yin Y, Duvall MR: The floral transcriptomes of four bamboo species (Bambusoideae; Poaceae). Submitted to *BMC Genomics*. January 2016. MS 34 pp. plus supplementary info.

Yamaguchi T, Lee DY, Miyao A, Hirochika H, An G, Hirano HY. 2006. Functional diversification of the two C-class MADS box genes OSMADS3 and OSMADS58 in Oryza sativa. The Plant Cell, 18(1): 15-28.

Yao SG, Ohmori S, Kimizu M, Yoshida H. 2008. Unequal genetic redundancy of rice PISTILLATA orthologs, OsMADS2 and OsMADS4, in lodicule and stamen development. Plant and Cell Physiology, 49(5): 853-857.

Zerbino, D.R. and Birney, E., 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research*, *18*(5), pp.821-829.

Zeng, C.X., Zhang, Y.X., Triplett, J.K., Yang, J.B. and Li, D.Z., 2010. Large multi-locus plastid phylogeny of the tribe Arundinarieae (Poaceae: Bambusoideae) reveals ten major lineages and low rate of molecular divergence. *Molecular Phylogenetics and Evolution*, *56*(2), pp.821-839.

Zhang, W., Chen, J., Yang, Y., Tang, Y., Shang, J. and Shen, B., 2011. A practical comparison of de novo genome assembly software tools for next-generation sequencing technologies. *PloS one*, *6*(3), p.e17915.

Zhang WP. 2000. Phylogeny of the Grass Family (Poaceae) from rpl16 Intron Sequence Data. Mol Phylogenet Evol. 15: 135–146

Zhang WP, Clark LG. 2000. Phylogeny and classification of the Bambusoideae (Poaceae). In: Jacobs SWL, Everett J (eds) Grass Syst Evol. CSIRO, Melbourne, p 35–42.0

Zhang XM, Zhao L, Larson-Rabin Z, Li DZ, Guo ZH. 2012. De novo sequencing and characterization of the floral transcriptome of *Dendrocalamus latiflorus* (Poaceae: Bambusoideae). PloS one, 7(8), e42082.

Zhang, Y.J., Ma, P.F. and Li, D.Z., 2011. High-throughput sequencing of six bamboo chloroplast genomes: phylogenetic implications for temperate woody bamboos (Poaceae: Bambusoideae). *PLoS One*, *6*(5), p.e20596.

Zhang LN, Zhang XZ, Zhang YX, Zeng CX, Ma PF, Zhao L, Guo ZH, Li DZ. 2014. Identification of putative orthologous genes for the phylogenetic reconstruction of temperate woody bamboos (Poaceae: Bambusoideae). Mol Ecol Resources, 14(5): 988-999.

Zizania aquatica [http://eol.org/pages/1114723/overview]