Title

The Streptochaeta Genome and the Evolution of the Grasses.

Authors
Seetharam, Arun
Yu, Yunqing
Bélanger, Sébastien
et al.

Publication Date
2021

DOI
10.3389/fpls.2021.710383

frontiers in Plant Science

ORIGINAL RESEARCH
published: 04 October 2021
doi: 10.3389/fpls.2021.710383

The *Streptochaeta* Genome and the Evolution of the Grasses

Arun S. Seetharam[1†], Yunqing Yu[2†], Sébastien Bélanger[2], Lynn G. Clark[1], Blake C. Meyers[2,3], Elizabeth A. Kellogg[2] and Matthew B. Hufford[1*]*

[1] Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, IA, United States, [2] Donald Danforth Plant Science Center, St. Louis, MO, United States, [3] Division of Plant Sciences, University of Missouri, Columbia, MO, United States

In this work, we sequenced and annotated the genome of *Streptochaeta angustifolia*, one of two genera in the grass subfamily Anomochlooideae, a lineage sister to all other grasses. The final assembly size is over 99% of the estimated genome size. We find good collinearity with the rice genome and have captured most of the gene space. *Streptochaeta* is similar to other grasses in the structure of its fruit (a caryopsis or grain) but has peculiar flowers and inflorescences that are distinct from those in the outgroups and in other grasses. To provide tools for investigations of floral structure, we analyzed two large families of transcription factors, AP2-like and R2R3 MYBs, that are known to control floral and spikelet development in rice and maize among other grasses. Many of these are also regulated by small RNAs. Structure of the gene trees showed that the well documented whole genome duplication at the origin of the grasses (ρ) occurred before the divergence of the Anomochlooideae lineage from the lineage leading to the rest of the grasses (the spikelet clade) and thus that the common ancestor of all grasses probably had two copies of the developmental genes. However, *Streptochaeta* (and by inference other members of Anomochlooideae) has lost one copy of many genes. The peculiar floral morphology of *Streptochaeta* may thus have derived from an ancestral plant that was morphologically similar to the spikelet-bearing grasses. We further identify 114 loci producing microRNAs and 89 loci generating phased, secondary siRNAs, classes of small RNAs known to be influential in transcriptional and post-transcriptional regulation of several plant functions.

Keywords: *Streptochaeta angustifolia*, grass evolution, spikelet, small RNA, APETALA2-like, R2R3 MYB

OPEN ACCESS

Edited by:
Jorge Duitama,
University of the Andes, Colombia

Reviewed by:
Domingos Cardoso,
Federal University of Bahia, Brazil
Diego Mauricio Riaño-Pachón,
University of São Paulo, Brazil

Correspondence:
Elizabeth A. Kellogg
ekellogg@danforthcenter.org
Matthew B. Hufford
mhufford@iastate.edu

† These authors have contributed equally to this work and share first authorship

Specialty section:
This article was submitted to Plant Systematics and Evolution, a section of the journal Frontiers in Plant Science

Received: *16 May 2021*
Accepted: *08 September 2021*
Published: *04 October 2021*

Citation:
Seetharam AS, Yu Y, Bélanger S, Clark LG, Meyers BC, Kellogg EA and Hufford MB (2021) The Streptochaeta Genome and the Evolution of the Grasses. Front. Plant Sci. 12:710383. doi: 10.3389/fpls.2021.710383

INTRODUCTION

The grasses (Poaceae) are arguably the most important plant family to humankind due to their agricultural and ecological significance. The diversity of grasses may not be immediately evident given their apparent morphological simplicity. However, the total number of described species in the family is 11,500+ (Soreng et al., 2017), and more continue to be discovered and described. Grasses are cosmopolitan in distribution, occurring on every continent. Estimates vary based on the definition of grassland, but, conservatively, grasses cover 30% of the Earth's land surface (White et al., 2000; Gibson, 2009). Grasses are obviously the major component of grasslands, but grass species also occur in deserts, savannas, forests (both temperate and tropical), sand dunes, salt marshes and freshwater systems, where they are often ecologically dominant (Lehmann et al., 2019).

The traits that have contributed to the long-term ecological success of the grasses have also allowed them to be opportunistic colonizers in disturbed areas and agricultural systems (Linder et al., 2018), where grasses are often the main crops, providing humanity with greater than 50% of its daily caloric intake (Sarwar, 2013). The adaptations and morphologies of the grasses that have led to ecological and agronomic dominance represent major innovations relative to ancestral species.

Monophyly of the grass family is unequivocally supported by molecular evidence, but grasses also exhibit several uniquely derived morphological or anatomical traits (Grass Phylogeny Working Group [GPWG], 2001; Kellogg, 2015; Leandro et al., 2018). These include the presence of arm cells and fusoid cells (or cavities) in the leaf mesophyll; the pollen wall with channels in the outer wall (intraexinous channels); the caryopsis fruit type; and a laterally positioned, highly differentiated embryo. The 30 or so species of the grass lineages represented by subfamilies Anomochlooideae, Pharoideae and Puelioideae, which are successive sisters to the remainder of the family, all inhabit tropical forest understories, and also share a combination of ancestral features including a herbaceous, perennial, rhizomatous habit; leaves with relatively broad, pseudopetiolate leaf blades; a highly bracteate inflorescence; six stamens in two whorls; pollen with a single pore surrounded by an annulus; a uniovulate gynoecium with three stigmas; compound starch granules in the endosperm; and the C_3 photosynthetic pathway (Grass Phylogeny Working Group [GPWG], 2001). The BOP (Bambusoideae, Oryzoideae, Pooideae) + PACMAD (Panicoideae, Aristidoideae, Chloridoideae, Micrairoideae, Arundinoideae, Danthonioideae) clade encompasses the remaining diversity of the family (Kellogg, 2015; **Figure 1A**). The majority of these lineages adapted to and diversified in open habitats, evolving relatively narrow leaves lacking both pseudopetioles and fusoid cells in the mesophyll, spikelets with an array of adaptations for dispersal, and flowers with three stamens and two stigmas. The annual habit evolved repeatedly in both the BOP and PACMAD clades, and the 24+ origins of C_4 photosynthesis occurred exclusively within the PACMAD clade (Grass Phylogeny Working Group II [GPWG II], 2012; Spriggs et al., 2014).

Anomochlooideae, a tiny clade of four species classified in two genera (*Anomochloa* and *Streptochaeta*), is sister to all other grasses (**Figure 1A**; Kellogg, 2015). Its phylogenetic position makes it of particular interest for studies of grass evolution and biology, particularly genome evolution. All grasses studied to date share a whole genome duplication (WGD), sometimes referred to as ρ, which is inferred to have occurred just before the origin of the grasses (Paterson et al., 2004; Wang et al., 2005; McKain et al., 2016). Not only are ancient duplicated regions found in the grass genomes studied to date, but the phylogenies of individual gene families often exhibit a doubly labeled pattern consistent with WGD (Rothfels, 2021). In this pattern we see, for example, a tree with the topology shown in **Figure 1B**, which points to a WGD before the divergence of all sequenced grasses, whereas a WGD after divergence of *Streptochaeta*, would result in the topology shown in **Figure 1C**. While there is some evidence from individual gene trees that the duplication precedes the divergence of *Streptochaeta* + *Anomochloa* (Preston and Kellogg, 2006; Preston et al., 2009; Christensen and Malcomber, 2012; Bartlett et al., 2016; McKain et al., 2016), data are sparse. Thus, defining the position of the grass WGD requires a whole genome sequence of a species of Anomochlooideae.

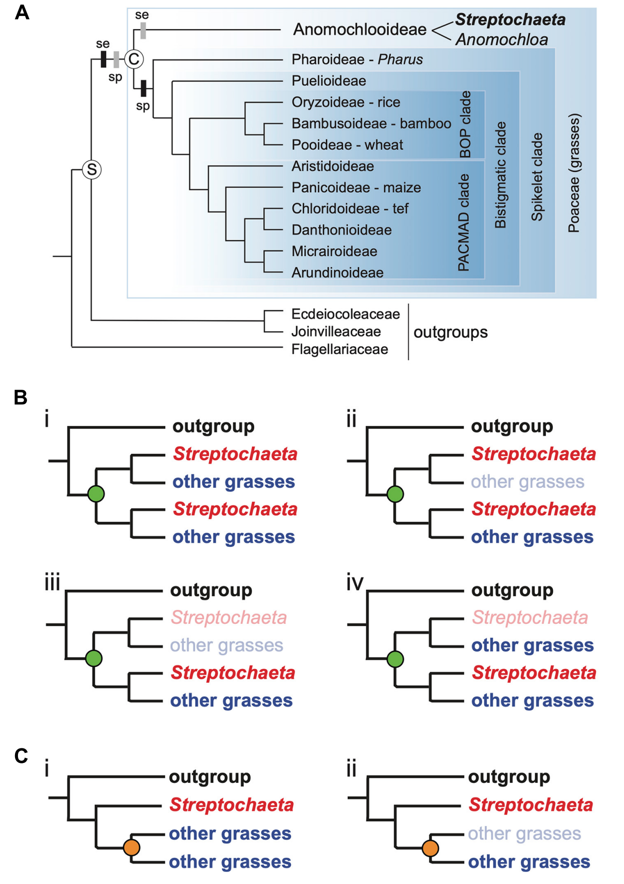

FIGURE 1 | Phylogenetic placement of *Streptochaeta*. **(A)** Phylogenetic tree depicting the BOP (Bambusoideae, Oryzoideae, Pooideae) + PACMAD (Panicoideae, Aristidoideae, Chloridoideae, Micrairoideae, Arundinoideae, Danthonioideae) clade and placement of focal organism *Streptochaeta* sister to the spikelet clade of grasses. Tree topology is well supported in most recent grass phylogenies (e.g., Saarela et al., 2018) except that in some analyses the relative positions of Aristidoideae and Panicoideae are switched. S, stem node of Poaceae; C, crown node. Black bars, stepwise model, in which spikelet equivalents (se) originate before the crown node and true spikelets (sp) originate afterward on the branch leading to the spikelet clade. Gray bars, loss model, in which spikelets (sp) originate before the crown node and then are modified to spikelet equivalents (se) afterward on the branch leading to Anomochlooideae. **(B,C)** Possible patterns of whole genome duplication (WGD) and gene loss. **(B)** WGD before the divergence of *Streptochaeta* assuming **(i)** no gene loss; **(ii)** loss of one clade of non-*Streptochaeta* grass paralogs soon after WGD; **(iii)** loss of all grass paralogs soon after WGD; **(iv)** loss of one *Streptochaeta* paralog soon after WGD. **(C)** WGD after divergence of *Streptochaeta*. **(i)** no gene loss; **(ii)** loss of one clade of non-*Streptochaeta* grass paralogs soon after WGD. Note that patterns **(Biii,Cii)** are indistinguishable.

Anomochlooideae is also in a key position for understanding the origins of the morphological innovations of the grass family

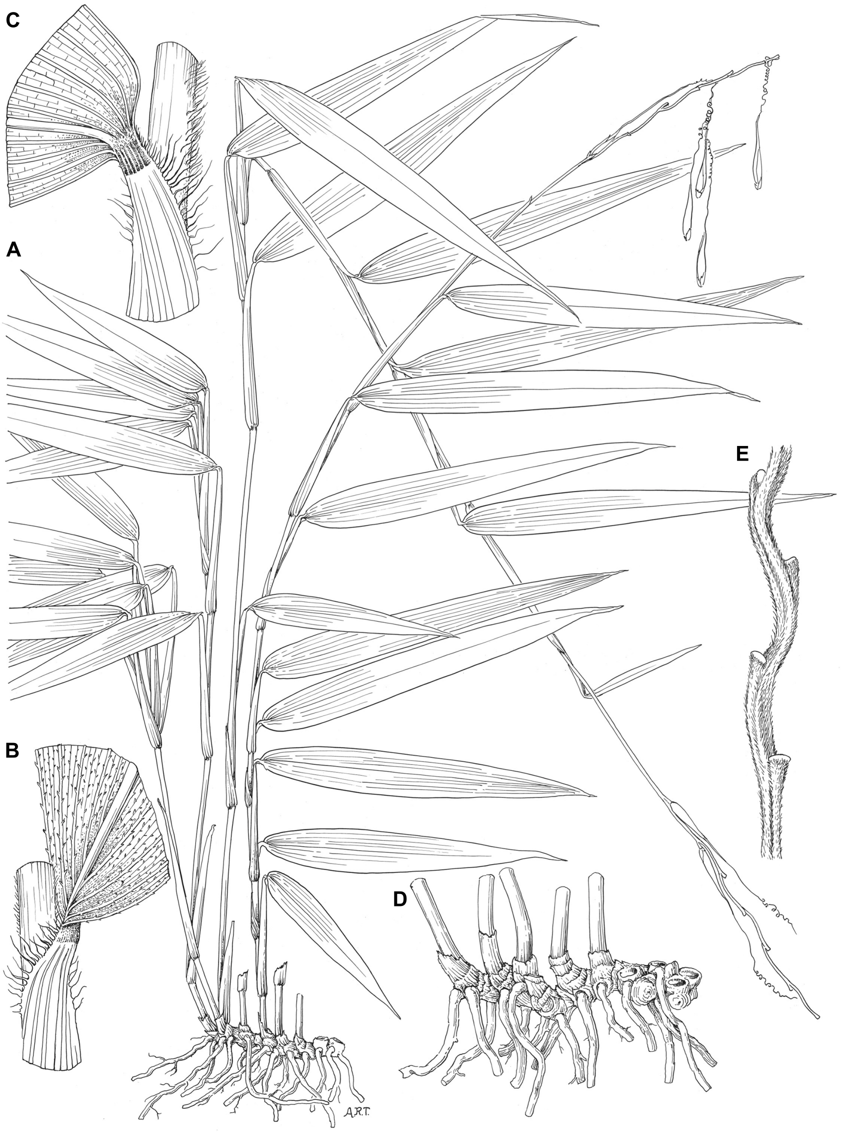

FIGURE 2 | *Streptochaeta angustifolia*. **(A)** Habit (×0.5). **(B)** Mid-region of leaf showing summit of sheath and upper surface of blade (×4.5). **(C)** Mid-region of leaf showing summit of sheath and lower surface of blade (×5). **(D)** Rhizome system with culm base (×1). **(E)** Portion of rachis enlarged (×1.5). All drawings based on Soderstrom and Sucre 1969 (US). Illustration by Alice R. Tangerini. Reprinted from Soderstrom (1981), originally **Figure 5**, p. 31, with permission from the Missouri Botanical Garden Press.

and in particular the evolution of the spikelet. Poaceae is sister to the clade of Ecdeiocoleaceae plus Joinvilleaceae and the three families in turn sister to Flagellariaceae (Magallón et al., 2015; Bouchenak-Khelladi et al., 2015; **Figure 1A**). The latter three families all have conventional 3-merous monocot flowers. In contrast, all grasses except Anomochlooideae bear their flowers in tiny clusters known as spikelets (little spikes) (Judziewicz et al., 1999; Grass Phylogeny Working Group [GPWG], 2001; Kellogg, 2015). Because the number, position, and structure of spikelets affect the total number of seeds produced by a plant, the genes controlling their development are a subject of continual research (e.g., Whipple, 2017; Huang et al., 2018; Li C. et al., 2019; Li Y. et al., 2019, to cite just a few). Unlike the rest of the Poaceae, the flowers in Anomochlooideae are borne in complex bracteate structures sometimes called "spikelet equivalents" (Soderstrom and Ellis, 1987; Judziewicz and Soderstrom, 1989; Judziewicz et al., 1999; **Figures 2**, **3**). These differ from both the conventional monocot flowers of the outgroups and the spikelets of the remainder of the grasses (i.e., the "spikelet clade," Sajo et al., 2008, 2012; Preston et al., 2009; Kellogg et al., 2013). In addition, the spikelet equivalents of *Anomochloa* and *Streptochaeta* also differ from each other such that it is difficult to establish unequivocal positional homologies among their parts. It is thus simplest to infer that the common ancestor of Ecdeiocoleaceae + Joinvilleaceae on the one hand and all grasses including Anomochlooideae on the other (i.e., the

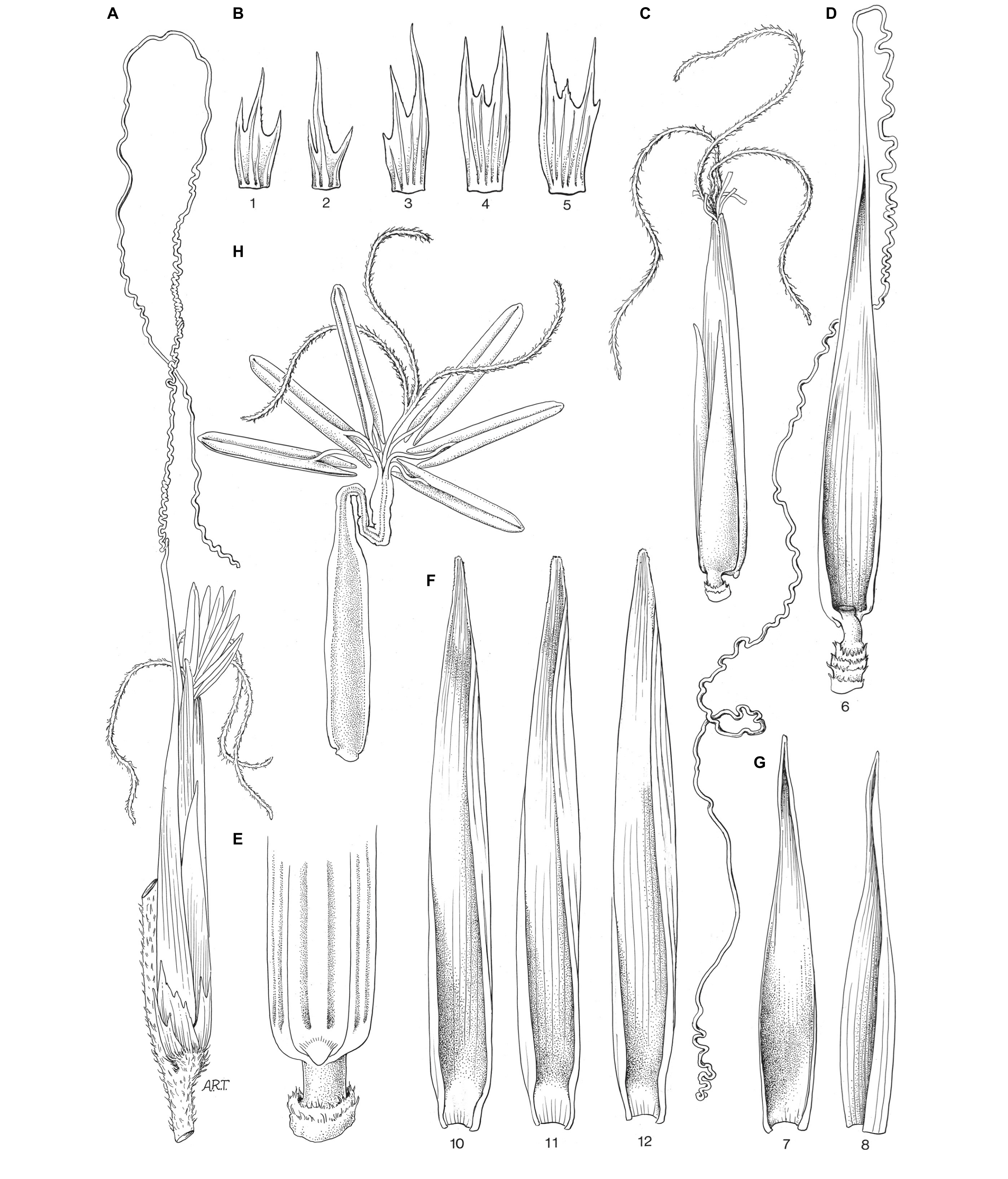

FIGURE 3 | *Streptochaeta angustifolia*. **(A)** Pseudospikelet (×4.5). **(B)** Series of bracts 1–5 from the base of the pseudospikelet (×6). **(C)** Pseudospikelet with basal bracts 1–5 removed and showing bracts 7 and 8, whose bases are overlapping (× 4.5). **(D)** Bract 6 with long coiled awn (×4.5). **(E)** Back portion of the base of bract 6 showing region where embryo exits at germination. **(F)** Bracts 10–12 (×6). **(G)** Bracts 7 and 8 (×6). Bract 9, which exists in other species, has not been found here. **(H)** Ovary with long style and three stigmas, surrounded by the thin, fused filaments of the 6 stamens (°4.5). All drawings based on Soderstrom and Sucre 1969 (US). Illustration by Alice R. Tangerini. Reprinted from Soderstrom (1981), originally **Figure 6** , p. 33, with permission from the Missouri Botanical Garden Press.

stem node of the grasses, S in **Figure 1A**) likely had standard 3-merous flowers but that sometime between the stem node and the crown node (C in **Figure 1A**) of Poaceae, floral or inflorescence development changed.

The phylogeny suggests at least two models for the inferred changes before and after the crown node of Poaceae. One possibility is a "stepwise" model (black bars in **Figure 1A**), in which a set of genetic changes before the crown node of the grasses led to floral units that were substantially different from those in other monocots and were similar to the spikelet equivalents of *Streptochaeta* and *Anomochloa*. After the crown node, floral development was further modified by a second set of changes that led to formation of true spikelets in the common ancestor of the spikelet clade. The alternative model (gray bars in **Figure 1A**), which is also consistent with the phylogeny, is a "loss model," in which all the genes and regulatory architecture needed for making spikelets originated before the crown node of Poaceae, but portions of that architecture were subsequently lost during the evolution of Anomochlooideae. Thus, the stepwise model implies that two successive sets of changes (one before and one after the crown node) were required for the origin of the grass spikelet, whereas the loss model implies a gain of spikelets followed by a loss; in this model the spikelet equivalents are highly modified or rearranged spikelets. Resolving these hypothetical models will help reveal both how the unique spikelet structure and the overall floral bauplan in grasses evolved.

Of the handful of species in the Anomochlooideae, *Streptochaeta angustifolia* (**Figures 2**, **3**) is the most easily grown from seed and an obvious candidate for ongoing functional genomic investigation. Hereafter in this paper, we will refer to *S. angustifolia* simply as *Streptochaeta*, and use it as a placeholder for the rest of the subfamily. We present a draft genome sequence for *Streptochaeta* that captures the gene-space of this species at high contiguity, and we use this genome to assess the position of the grass WGD. Genes and small RNAs (sRNAs) are annotated. Because of the distinct floral morphology of *Streptochaeta*, we also investigate the molecular evolution of two major transcription factor families, APETALA2 (AP2)-like and R2R3 MYB, which are known to control floral and spikelet structure in other grasses and are regulated by sRNAs.

MATERIALS AND METHODS

Input Data

Streptochaeta angustifolia is native and restricted to the Atlantic Forest of Brazil, although other species of *Streptochaeta* can be found as far north as southern Mexico. The reference plant for this project was collected in Brazil by Thomas Soderstrom of the Smithsonian Institution in 1980 of the Smithsonian Institution, though the precise collection location is unknown. The plant has been propagated by division and single seed descent, first at the Smithsonian and more recently at Iowa State University and at the Donald Danforth Plant Science Center in St. Louis, MO. The voucher *Clark 1304* (deposited at Ada Hayden Herbarium, ISC) represents the plant from which DNA was extracted for the initial molecular phylogenetic studies of this genus. Either this plant or one of its descendants was used for this project, based on the same voucher.

Streptochaeta leaf tissue was harvested and used to estimate genome size at the Flow Cytometry Facility at Iowa State University. DNA was then isolated using Qiagen DNeasy plant kits. Three Illumina libraries (paired end and 9- and 11-kb mate pair) were generated from these isolations at the Iowa State University (ISU) DNA Facility. One lane of 150 bp paired-end HiSeq sequencing (insert size of 180 bp) and one lane of 150 bp mate-pair HiSeq sequencing (9- and 11-kb libraries pooled) were generated, also at the ISU DNA Facility (**Supplementary Table 1**). Additionally, for the purpose of contig scaffolding, Bionano libraries were prepared by first isolating high molecular weight DNA using the Bionano Prep^{TM} Plant DNA Isolation Kit followed by sequencing using the Irys system.

Genome Assembly

We used MaSuRCA v2.21 (Zimin et al., 2013) to generate a draft genome of *Streptochaeta*. The MaSuRCA assembler includes error correction and quality filtering, generation of super reads, super read assembly, and gap closing to generate more complete and larger scaffolds. Briefly, the config file was edited to include both paired-end and mate-pair library data for *Streptochaeta.* The JF_SIZE parameter was adjusted to 20,000,000,000 to accommodate the large input file size, and NUM_THREADS was set to 128. All other parameters in the config file were left as default. The assembly was executed by first generating the assemble.sh script using the config file and submitting to a high-memory node using the PBS job scheduler. For generating the Bionano-optical-map-based hybrid assembly, we used Bionano Hybrid Scaffold (v1.0). This program uses the alignment of *in silico*-generated maps (from input contigs) to the consensus optical map (Bionano) to output genome maps. The genome maps are then aligned back to the original *in silico* maps to output fasta-formatted hybrid scaffolds (called Super Scaffolds). The full list of options used for running the alignment and to generate the hybrid scaffolds are provided in the associated GitHub repository (files: optArguments_medium.xml and hybridScaffold_config_aggressive.xml, respectively). All scripts for assembly and downstream analysis are available at: https://github.com/HuffordLab/streptochaeta.

Assembly Evaluation and Post-processing

The Bionano assembly was screened for haplotigs, and additional gaps were filled using Redundans v0.13a (Pryszcz and Gabaldón, 2016). Briefly, the scaffolds were mapped to themselves using the LAST v719 alignment program (Kielbasa et al., 2011) and any scaffold that completely overlapped a longer scaffold with more than 80% identity was considered redundant and excluded from the final assembly. Additionally, short read data were aligned back to the hybrid assembly and GapCloser v1.12 from SOAPdenovo2 (Luo et al., 2012) and SSPACE v3.0 (Boetzer et al., 2011) were run in multiple iterations to fill gaps. The final reduced, gap-filled assembly was screened for contamination, using Blobtools v0.9.19

(Laetsch and Blaxter, 2017), and any scaffolds that matched bacterial genomes were removed. The assembly completeness was then evaluated using BUSCO v3.0.2 (Simão et al., 2015) with the liliopsida_odb10 profile and standard assemblathon metrics. We used Merqury (v1.3; Rhie et al., 2020) to estimate the frequency of consensus errors (consensus quality or QV) and k-mer completeness.

To annotate the repeats in the genome, we used EDTA v1.8.3 (Ou et al., 2019) with default options except for –species, which was set to "others." The obtained TE library was then used for masking the genome for synteny analyses. Assembly quality of the repeat space was assessed based on the LTR Assembly Index (LAI; Ou et al., 2018), which was computed using ltr_retriever v2.9.0 (Ou and Jiang, 2018) and the EDTA-generated LTR list.

Gene Prediction and Annotation

Gene prediction was carried out using a comprehensive method combining *ab initio* predictions (from BRAKER; Hoff et al., 2019) with direct evidence (inferred from transcript assemblies) using the BIND strategy (Seetharam et al., 2019 and citations therein). Briefly, RNA-Seq data were mapped to the genome using a STAR (v2.5.3a)-indexed genome and an iterative two-pass approach under default options in order to generate BAM files. BAM files were used as input for multiple transcript assembly programs (Class2 v2.1.7, Cufflinks v2.2.1, Stringtie v2.1.4 and Strawberry v1.1.2) to assemble transcripts. Redundant assemblies were collapsed and the best transcript for each locus was picked using Mikado (2.0rc2) by filling in the missing portions of the ORF using TransDecoder (v5.5.0) and homology as informed by the BLASTX (v2.10.1+) results to the SwissProtDB. Splice junctions were also refined using Portcullis (v1.2.1) in order to identify isoforms and to correct misassembled transcripts. Both *ab initio* and the direct evidence predictions were analyzed with TESorter (Zhang et al., 2019) to identify and remove any TE-containing genes and with phylostratr (v0.20; Arendsee et al., 2019) to identify orphan genes (i.e., species-specific genes). As *ab initio* predictions of young genes can be unreliable (Seetharam et al., 2019), these were excluded. Finally, redundant copies of genes between direct evidence and *ab initio* predictions were identified and removed using Mikado compare (2.0rc2; Venturini et al., 2018) and merging was performed locus by locus, incorporating additional isoforms when necessary. The complete decision table for merging is provided in **Supplementary Table 2**. After the final merge, phylostratr was run again on the annotations to classify genes based on their age.

Functional annotation was performed based on homology of the predicted peptides to the curated SwissProt/UniProt set (UniProt Consortium, 2021) as determined by BLAST v2.10.1+ (Edgar, 2010). InterProScan v5.48-83.0 was further used to find sequence matches against multiple protein signature databases.

Synteny

Synteny of CDS sequences for *Streptochaeta* was determined using CoGe (Lyons and Freeling, 2008), against the genomes Brachypodium (International Brachypodium Initiative [IBI], 2010), *Oryza sativa* (Ouyang et al., 2007), and *Setaria viridis* (Mamidi et al., 2020). SynMap2 (Haug-Baltzell et al., 2017) was employed to identify syntenic regions across these genomes. Dot plots and chain files generated by SynMap2 under default options were used for presence–absence analysis. We also performed repeat-masked whole genome alignments using minimap2 (Li, 2018) following the Bioinformatics Workbook methods[1].

Identification of APETALA2-Like and R2R3 MYB Proteins in Selected Monocots

A BLAST database was built using seven grass species including *Streptochaeta* and two outgroup monocots. Protein and CDS sequences of the following species were retrieved from Phytozome 13.0: *Ananas comosus* (Acomosus_321_v3), *Brachypodium distachyon* (Bdistachyon_556_v3.2), *Oryza sativa* (Osativa_323_v7.0), *Spirodela polyrhiza* (Spolyrhiza_290_v2), *Setaria viridis* (Sviridis_500_v2.1), and *Zea mays* (Zmays_493_APGv4). Sequences of *Eragrostis tef* were retrieved from CoGe (id50954) (VanBuren et al., 2020). Sequences of *Triticum aestivum* were retrieved from Ensembl Plant r46 (Triticum_aestivum.IWGSCv1) (**Supplementary Table 3**).

AP2 and MYB proteins were identified using BLASTP and hmmscan (HMMER 3.1b2[2]) in an iterative manner. Specifically, 18 *Arabidopsis* AP2-like proteins (Kim et al., 2006) were used as an initial query in a blastp search with an *E*-value threshold of 1e-10. The resulting protein sequences were filtered based on the presence of an AP2 domain using hmmscan with an *E*-value threshold of 1e-3 and domain *E*-value threshold of 0.1. The filtered sequences were used as the query for the next round of blastp and hmmscan until the maximal number of sequences was retrieved. For MYB proteins, Interpro MYB domain (IPR017930) was used to retrieve rice MYBs using *Oryza sativa* Japonica Group genes (IRGSP-1.0) as the database on Gramene Biomart[3]. The number of MYB domains was counted by searching for "Myb_DNA-bind" in the output of hmmscan, and 82 proteins with two MYB domains were used as the initial query. Iterative blastp and hmmscan were performed in the same manner as for AP2 except using a domain *E*-value threshold of 1e-3.

The number of AP2 or MYB domains was again counted in the final set of sequences in the hmmscan output, and proteins with more than one AP2 domain or two MYB domains were treated as AP2-like or R2R3 MYB, respectively. To ensure that no orthologous proteins were missed due to poor annotation in the AP2 or MYB domain, we performed another round of BLASTP searches, and kept only the best hits. These sequences were also included in the construction of the phylogenetic trees.

Construction and Rooting of Phylogenetic Trees

Protein sequences were aligned using MAFFT v7.245 (Katoh and Standley, 2013) with default parameters. The corresponding

[1] https://bioinformaticsworkbook.org/dataWrangling/genome-dotplots.html

[2] http://hmmer.org/

[3] http://ensembl.gramene.org/biomart/martview/

coding sequence alignment was converted using PAL2NAL v14 (Suyama et al., 2006) and used for subsequent tree construction. For *AP2*-like genes, the full-length coding sequence alignment was used. For MYB, due to poor alignment outside of the MYB domain, trimAl v1.2 (Capella-Gutiérrez et al., 2009) was used to remove gaps and non-conserved nucleotides with a gap threshold (–gt) of 0.75 and percentage alignment conservation threshold (-con) of 30. A maximum likelihood (ML) tree was constructed using IQ-TREE v1.6.12 (Minh et al., 2020) with default settings. Sequences that resulted in long branches in the tree were manually removed, and the remaining sequences were used for the final tree construction. Visual formatting of the tree was performed using Interactive Tree Of Life (iTOL) v4 (Letunic and Bork, 2019).

The ML tree for *AP2*-like genes was rooted at the branch between the euAP2 and AINTEGUMENTA (ANT) genes, following (Kim et al., 2006). The tree of R2R3 MYBs was rooted with the CDC5 clade (Jiang and Rao, 2020). Only subclades with bootstrap values larger than 80 at the node of Streptochaeta were considered for subsequent analysis.

To facilitate discussion, we named each subclade either by a previously assigned gene name within the subclade, or the gene sub-family name with a specific number.

RNA Isolation, Library Construction, and Sequencing

To annotate microRNAs (miRNAs) present in the *Streptochaeta* genome, we (i) sequenced sRNAs from leaf, anther and pistil tissues, (ii) compared miRNAs present in anthers to those of three other representative monocots (rice, maize, and asparagus), and (iii) validated gene targets of these miRNAs.

We collected tissues from leaf and pistil as well as 1.5, 3, and 4 mm anthers. Samples were immediately frozen in liquid nitrogen and kept at −80°C prior to RNA isolation. Total RNA was isolated using the PureLink Plant RNA Reagent (Thermo Fisher Scientific, Waltham, MA, United States). sRNA libraries were published previously (Patel et al., 2021). RNA sequencing libraries were prepared from the same material using the Illumina TruSeq stranded RNA-seq preparation kit (Illumina Inc., United States) following manufacturer's instructions. Parallel analysis of RNA ends (PARE) libraries were prepared from a total of 20 μg of total RNA following the method described by Zhai et al. (2014). For all types of libraries, single-end sequencing was performed on an Illumina HiSeq 2000 instrument (Illumina Inc., United States) at the University of Delaware DNA Sequencing and Genotyping Center.

Bioinformatic Analysis of Small RNA Data

Using cutadapt v2.9 (Martin, 2011), sRNA-seq reads were pre-processed to remove adapters (**Supplementary Table 4**), and we discarded reads shorter than 15 nt. The resulting 'clean' reads were mapped to the *Streptochaeta* genome using ShortStack v3.8.5 (Johnson et al., 2016) with the following parameters: -mismatches 0, -bowtie m 50, -mmap u, -dicermin 19, -dicermax 25, and -mincov 0.5 transcripts per million (TPM). Results generated by ShortStack were filtered to keep only clusters having a predominant RNA size between 20 and 24 nucleotides, inclusively. We then annotated categories of miRNAs and phased small interfering RNAs (phasiRNAs).

First, sRNA reads representative of each cluster were aligned to the monocot-related miRNAs listed in miRBase release 22 (Kozomara and Griffiths-Jones, 2014; Kozomara et al., 2019) using NCBI BLASTN v2.9.0^{+} (Camacho et al., 2009) with the following parameters: -strand both, -task blastn-short, -perc identity 75, -no greedy and -ungapped. Homology hits were filtered and sRNA reads were considered as known miRNA based on the following criteria: (i) no more than four mismatches and (ii) no more than 2-nt extension or reduction at the 5′ end or 3′ end. Known miRNAs were summarized by family. Small RNA reads with no homology to known miRNAs were annotated as novel miRNAs using the de novo miRNA annotation performed by ShortStack. The secondary structure of new miRNA precursor sequences was drawn using the RNAfold v2.1.9 program (Lorenz et al., 2011). Candidate novel miRNAs were manually inspected, and only those meeting published criteria for plant miRNA annotations (Axtell and Meyers, 2018) were retained for subsequent analyses. Then, the remaining sRNA clusters were analyzed to identify phasiRNAs based on ShortStack analysis reports. sRNA clusters having a "Phase Score" >30 were considered as true positive phasiRNAs. Genomic regions corresponding to these phasiRNAs were considered as PHAS loci and grouped in categories of 21- and 24-PHAS loci referring to the length of phasiRNAs derived from these loci. Other sRNA without miRNA or phasiRNA signatures were not considered for analysis or interpretation in this study.

To compare sRNAs accumulating in *Streptochaeta* anthers with other monocots, we analyzed sRNA samples of *Asparagus officinalis*, *Oryza sativa* and *Zea mays* anthers. The GEO accession numbers for those datasets are detailed in **Supplementary Table 3**. We analyzed these data as described for the *Streptochaeta* sRNA-seq data.

We used the upSetR package (Lex et al., 2014; Conway et al., 2017; UpSetR, 2021) to visualize the overlap of miRNA loci annotated in *Streptochaeta*, compared to other species.

Bioinformatic Analysis of Parallel Analysis of RNA Ends Data

We analyzed the PARE data to identify and validate miRNA-target pairs in anther, pistil, and leaf of *Streptochaeta* tissues. Using cutadapt v2.9, PARE reads were pre-processed to remove adapters (**Supplementary Table 4**) and reads shorter than 15 nt were discarded. Then, we used PAREsnip2 (Thody et al., 2018) to predict all miRNA-target pairs and to validate the effective miRNA-guided cleavage site using PARE reads. We ran PAREsnip2 with default parameters using Fahlgren and Carrington targeting rules (Fahlgren and Carrington, 2010). We considered only targets in categories 0, 1, and 2 for downstream analysis. We used the EMBL-EBI HMMER program v3.3 (Potter et al., 2018) to annotate the function of miRNA target genes using the phmmer function with the SwissProt database.

Prediction of MicroRNA Binding Sites

Mature miR172 and miR159 sequences from all available monocots were obtained from miRBase (Kozomara et al., 2019). miRNA target sites in *AP2*-like and *R2R3 MYB* transcripts were predicted on the web server TAPIR (Bonnet et al., 2010) with their default settings (score = 4 and free energy ratio = 0.7).

RESULTS

Genome Assembly

Flow cytometry estimated the 1C DNA content for *Streptochaeta* to be 1.80 and 1.83 pg, which, when converted to base pairs, yields a genome size of approximately 1.77 Gb. Paired-end reads with a fragment size of 250 bp were generated at approximately 25.7x genomic coverage, while the mate-pair libraries with 9- and 11-kb insert size collectively provided 22.6x coverage. Based on k-mer analysis with the program Jellyfish (Marçais and Kingsford, 2011), we estimated the repeat content for the *Streptochaeta* genome to be approximately 51%. The MaSuRCA assembly algorithm generated an assembly size at 99.8% of the estimated genome size, suggesting that much of the genome, including repetitive regions was successfully assembled. The assembler generated a total of 22,591 scaffolds, with an N50 of 2.4 Mb and an L50 of 170.

The Bionano data produced an optical map near the expected genome size (1.74 Gb) with an N50 of 824 kb. Through scaffolding with the optical map and collapsing with Redundans software, the total number of scaffolds dropped to 17,040, improving the N50 to 2.6 Mb and the L50 to 161. A total of 79,165 contigs were provided as input for Redundans for scaffold reduction (total size 1,898 Mbp). With eight iterations of haplotype collapsing, the total size reduced to 1,796 Mbp. Additional rounds of gap-filling using GapCloser reduced the total size of gaps (Ns) from 210.13 to 76.33 Mbp. The improvement in the N50/N90 values with each iteration is provided in **Supplementary Table 5**.

The final assembly included a total of 3,010 out of 3,278 possible complete Liliopsida BUSCOs (91.8%). Of these 2,767 (84.4% of the total) were present as a complete single copy. Only 158 BUSCOs were missing entirely with another 110 present as fragmented genes. The LAI (LTR Assembly Index) score, which assesses the contiguity of the assembled LTR retrotransposons, was 9.02, which is somewhat higher than most short-read-based assemblies (Ou et al., 2018), perhaps due to the relatively low repeat content of the *Streptochaeta* genome and the use of mate-pair sequencing libraries. Merqury's log-scaled probability of error in the consensus base calls indicated a high QV of 47.2146 (or probability of finding an incorrect base in this assembly was 1.89908e-05). The K-mer completeness measured by Merqury indicated 95.96% of reliable k-mers that occurred in the raw reads were also in the genome assembly, suggesting most of the reads were incorporated in the genome assembly. Dot plots of *Streptochaeta* contigs aligned to rice revealed substantial colinearity (**Supplementary Figure 1**).

BlobTools (v0.9.19) (Laetsch and Blaxter, 2017) detected over 95% of the scaffolds (1742 Mbp) belonging to the Streptophyta clade out of the 1,797 Mbp of assigned scaffolds (GC mean: 0.54). Approximately 2% of the scaffolds mapped to the Actinobacteria (36.3 Mbp, GC mean: 0.72) and ~0.5% of scaffolds to Chordata (9 Mbp, GC mean: 0.48). Scaffolds assigned to additional clades by BlobTools collectively comprise ~1.46 Mbp and the remaining 8.47 Mbp of scaffolds lacked any hits to the database. All bacterial, fungal and vertebrate scaffolds were purged from the assembly.

Gene Prediction and Annotation

Direct Evidence Predictions

More than 79% of the total RNAseq reads mapped uniquely to the *Streptochaeta* genome with < 7% multi-mapped reads. Paired-end reads mapped (uniquely) at a higher rate (88.59%) than the single-end RNAseq (70.38%) reads. Genome-guided transcript assemblers produced varying numbers of transcripts across single-end (SE) and paired-end (PE) data as well as various assemblers. Cufflinks produced the highest number of transcripts (SE: 65,552; PE:66,069), followed by StringTie (SE: 65,495, PE: 48,111), and Strawberry (SE:68,812; PE:43,882). Class2 generated fewer transcripts overall (PE: 43,966; SE: 13,173). The best transcript for each locus was picked by Mikado from the transcript assemblies based on its completeness, homology, and accuracy of splice sites. Non-coding (due to lack of ORFs) or redundant transcripts were removed to generate 28,063 gene models (41,857 transcripts). Mikado also identified 19,135 non-coding genes within the transcript assemblies. Further filtering for transposable-element-containing genes and genes with low expression reduced the total number of evidence-based predictions to 27,082 genes (40,865 transcripts).

Ab initio Predictions

BRAKER, with inputs including predicted proteins from the direct evidence method (as a gff3 file produced by aligning proteins to a hard-masked *Streptochaeta* genome) and the mapped RNA-Seq reads (as a hints file using the bam file), produced a total of 611,013 transcripts on a soft-masked genome. This was then subjected to filtering to remove TE containing genes (244,706 gene models) as well as genes only found in *Streptochaeta* (466,839 gene models). After removing both of these classes of genes, which overlapped to an extent, the total number of *ab initio* predictions dropped to 40,921 genes (44,013 transcripts).

BIND (Merging BRAKER Predictions With Directly Inferred Genes)

After comparing BRAKER and direct evidence predictions with Mikado compare: 9,617 transcripts were exactly identical and direct evidence predictions were retained; 3,263 transcripts from Mikado were considered incomplete and were replaced with BRAKER models; 13,360 BRAKER models were considered incomplete and replaced with direct evidence transcripts; 1,884 predictions were adjacent but non-overlapping, and 17,894 predictions were BRAKER-specific and were retained in the final merged predictions. The final gene set included a total of 44,980 genes (58,917 transcripts).

Functional Annotation

Functional annotation was informed by homology to the curated proteins in SwissProt and resulted in the assignment of putative functions for 38,955 transcripts (10,556 BRAKER predictions, and 28,399 direct evidence predictions). Of the unassigned transcripts, 41 predictions had pfam domain matches, and 16,918 transcripts had an interproscan hit. Only 3,068 transcripts contained no additional information in the final GFF3 file.

Phylostrata

All gene models were classified based on their presumed age. More than 8% of the total genes (3,742) were specific to the *Streptochaeta* genus and more than 15% (6,930) were Poaceae specific. Nineteen percent (8,494) of genes' origins could be traced back to cellular organisms and 15% (6,708) to Eukaryotic genes. The distribution of genes based on strata and annotation method is provided in **Supplementary Table 6**.

Transposable Element Annotation

The repeat annotation performed by the EDTA package comprised 66.82% of the genome, the bulk of which was LTR class elements (42.9% in total; Gypsy: 28.16%, Copia: 8.9%, rest: 5.84%), followed by DNA repeats (23.39% in total; DTC-type: 13.65, DTM-type: 5.78%, rest: 3.96%), and MITE class repeats (all types 0.54%).

Molecular Evolution of APETALA2-Like and R2R3 MYB Transcription Factors

APETALA2-Like

The *AP2*-like genes were divided into euAP2 and ANT clades, as expected from previous work (Kim et al., 2006). The euAP2 lineage has conserved microRNA172 binding sequences except for a few genes in outgroups, one gene in *Eragrostis tef* and one in *Zea mays* (**Figure 4** and **Supplementary Figure 2**).

Streptochaeta orthologs are present in most of the subclades, except *IDS1*/*Q*, *ANT5*, *BBM4*, *WRI3* and *basalANT1*, in which the *Streptochaeta* copy is lost (**Figure 4** and **Supplementary Figure 2**). The two most common patterns within each subclade are [O,(S,G)] (O, outgroup; S, *Streptochaeta*; G, other grasses) including *SHAT1*, *ANT1*, *ANT3*, *ANT4*, *BBM1*, *ANT7*, *ANT8,* and *ANT9*, and (S,G) (implying that the outgroup sequence is lost or was not retrieved by our search) including *BBM3*, *WRI2* and *WRI4* (**Supplementary Table 7**). These patterns imply that most grass-duplicated *AP2*-like genes were lost (i.e., the individual subclades were returned to single copy) soon after the grass duplication. Some subclades contain two *Streptochaeta* sequences and one copy in other grasses. These *Streptochaeta* sequences are either sisters to each other with the *Streptochaeta* clade sister to the other grasses [O,((S1,S2),G)] (*RSR1*) (**Figure 4**, **Supplementary Figure 2**, and **Supplementary Table 7**), or successive sisters to a clade of grass sequences [O,(S1,(S2,G))] (*WRI1*) (**Figure 4**, **Supplementary Figure 2**, and **Supplementary Table 7**).

In the paired subclades of *IDS1*/*Q-SNB*/*SID1*, *ANT5–ANT6*, *BBM4–BBM2* and *basalANT1-basalANT2*, the grass-duplicated gene pairs were retained, and were also in syntenic regions based on a syntelog search of the *Brachypodium distachyon*, *Oryza sativa* or *Setaria viridis* genomes (**Figure 5**). Interestingly, in these subclade pairs, the *Streptochaeta* orthologs are always sister to one member of the syntenic gene pair but not the other. Two subclade pairs support a ρ position before the divergence of *Streptochaeta*, including *BBM4-BBM2* with a pattern of [G1,(S,G2)] (**Figure 5B**) and *ANT5-ANT6* with a pattern of [G1,((S1,S2),G2)] (**Figure 5E**). In subclade pairs of *IDS1*/*Q-SNB*/*SID1* and *basalANT1-basalANT2*, two *Streptochaeta* sequences are successive sisters to one of the grass subclade pairs, forming tree topologies of [G1,(S1,(S2,G2))] and [O,(G1,(S1,(S2,G2)))], respectively (**Figure 4**, **Supplementary Figure 2**, and **Supplementary Table 7**). These two cases do not fit with a simple history involving ρ either before or after the divergence of *Streptochaeta*, and thus indicate a more complex evolutionary history.

R2R3 MYB

As in the *AP2*-like tree, the most common tree topology within each subclade is [O,(S,G)], found in 16 individual subclades, followed by (S,G) in 10 subclades. We also found 16 subclades with other tree topologies either without or with one or two *Streptochaeta* sequences and one copy of the other grass sequences, including (O,G) (*MYB48*), [O,((S1,S2),G)] (*MYB17*, *MYB21*, *GAMYBL2*, *MYB29* and *GAMYBL1*), [(S1,S2),G] (*MYB78* and *MYB92*), [O,(S1,(S2, G)))], [S1,(S2,G)] (*MYB56*) and [(O,S),G] (*MYB47* and *MYB83*) (**Supplementary Table 7**). Conversely, we also found that 20 subclade pairs retained the grass duplicated gene pairs, although their tree topologies vary based on the position of *Streptochaeta* and outgroups. Among these, 15 subclade pairs are also found to be syntenic, including *MYB1–MYB2*, *MYB6–MYB7*, *MYB35–MYB36*, *MYB42–MYB43*, *MYB49–MYB50*, *MYB51–MYB52*, *MYB53–MYB54*, *MYB62–MYB63*, *MYB65–MYB66*, *SWAM1–SWAM2*, *MYB75–MYB76*, *MYB86–MYB87*, *MYB93–MYB94*, *MYB103–MYB104,* and *MYB105-FDL1* (**Figures 5**, **6**, **Supplementary Figure 3**, and **Supplementary Table 7**). Together, these results indicate that a subset of grass MYB clades have expanded due to the grass WGD.

Among the subclade pairs that retain both grass sequences, we found one subclade pair, *MYB53-MYB54* with tree topology of [O,(S1,S2),(G1,G2)], that supports ρ having occurred after the divergence of *Streptochaeta* (**Figure 5F**). Conversely, we found 10 subclades supporting a ρ position before the divergence of *Streptochaeta*. The subclade *MYB93–MYB94* includes three *Streptochaeta* sequences, one sister to one of the grass clades and the other two sister to each other and sister to the other grass clade, forming a tree topology of [O,((S1,G1),((S2,S3),G2))] (**Figure 5A**). In the other nine subclade pairs, one or two *Streptochaeta* sequences are sister to one of the grass syntenic gene pairs but not the other (**Figures 5B–E**). In subclade pairs *MYB86–MYB87* and *MYB34–MYB36*, one *Streptochaeta* sequence is sister to one of the grass clades, showing [G1,(S,G2)] and [O,(G1,(S,G2))], respectively (**Figures 5B,C**). We observed more subclades with two sequences of *Streptochaeta*, either showing [O,(G1,((S1,S2),G2))] in *MYB6–MYB7* and *SWAM1* and *SWAM2*, or [G1,((S1,S2),G2)] in *MYB42–MYB43*, *MYB51–MYB52*, *MYB65–MYB66*, *MYB75–MYB76,* and *MYB105-FDL1*.

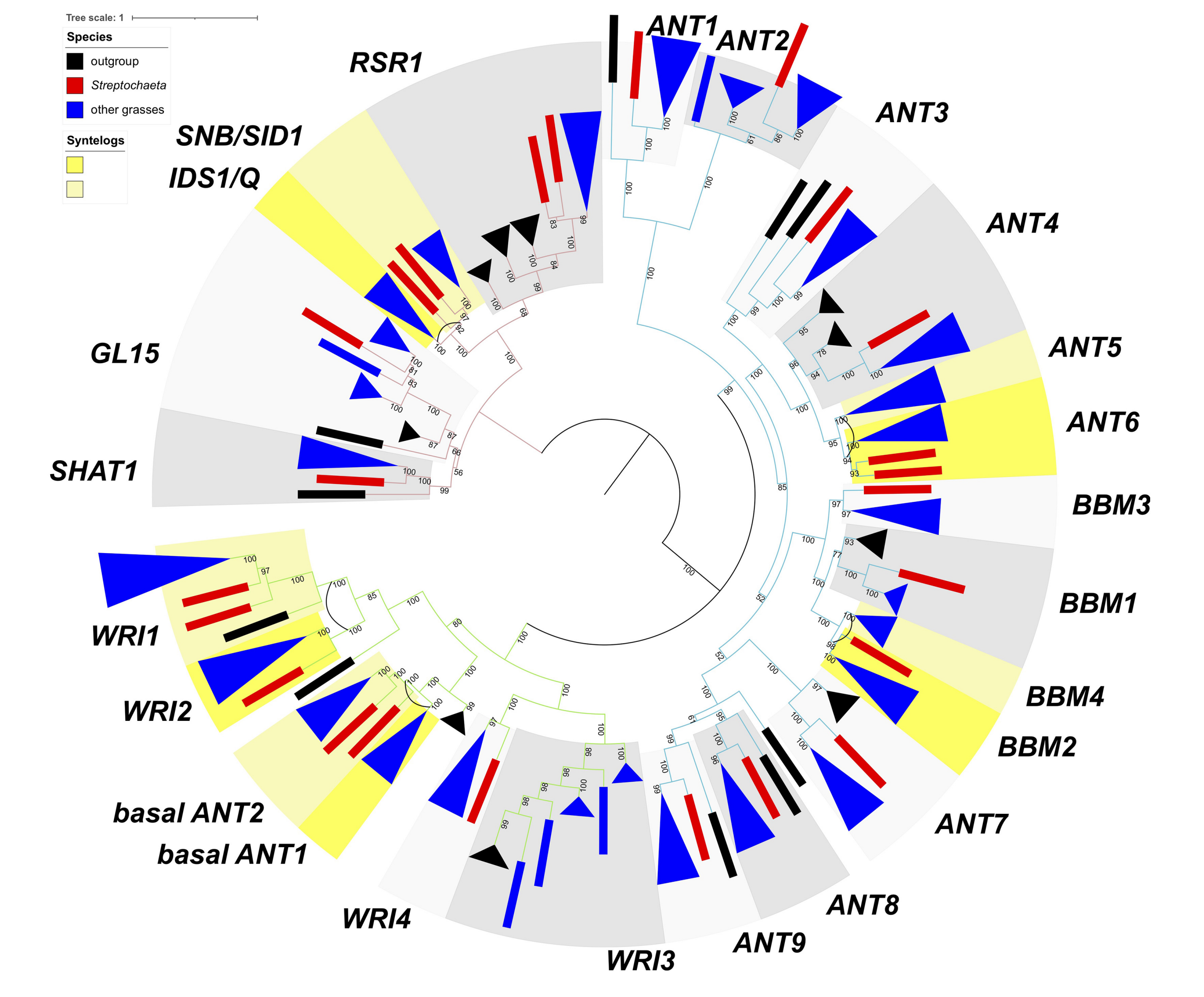

FIGURE 4 | Maximum likelihood tree of *AP2*-like genes. Numbers on branches indicate maximum likelihood bootstrap values. A single gene is denoted by a rectangle, and collapsed branches are denoted by triangles. Each subclade is shaded in two gray colors and named either by known genes within the subclade or subfamily name with a number. Subclades with syntenic genes in *Brachypodium*, *Oryza* or *Setaria* are shaded in two colors of yellow, and syntenic pairs are connected by an arc. Outgroup, *Streptochaeta* and other grasses are shown in black, red, and blue colors.

A few subclade pairs have tree topologies that do not support a ρ position either before or after the divergence of *Streptochaeta*, including [O,(S1,(S2,(G1,G2)))] (*MYB1–MYB2* and *MYB62–MYB63*), [S1,(G1,(S2,G2))] (*MYB22–MYB23*) and [(O,S),(G1,G2)] (*MYB11–MYB12*) (**Supplementary Table 7**). In other cases, the *Streptochaeta* ortholog is either lost, or positioned within the grass clades (**Supplementary Table 7**). This may either indicate a complex evolutionary history within the *Streptochaeta lineage*, or may be an artifact due to the distant outgroups used here and/or poor annotation of some sequences.

Taken together, both the *AP2*-like and *R2R3 MYB* trees support the inference of ρ before the divergence of *Streptochaeta* (12 subclades) over ρ after the divergence of *Streptochaeta* (1 subclade) (**Figure 5**), consistent with previous findings (McKain et al., 2016). In addition, our study suggests that *Streptochaeta* has often lost one of the syntenic paralogs and sometimes has its own duplicated gene pairs.

Annotation of MicroRNAs and Validation of Their Targets

In total, 185.3 million (M) sRNA reads were generated (115.6 M, 33.0 M, and 36.7 M reads for anther, pistil, and leaf tissues, respectively) from five sRNA libraries. We annotated 114 miRNA loci, of which 98 were homologous to 32 known miRNA families and 16 met strict annotation criteria for novel miRNAs (**Supplementary Tables 8–10**). Most miRNAs from these loci (85; 90.4%) accumulated in all three tissues (**Figure 7**). However, a sub-group (8 miRNAs; 7.0%) of miRNAs was abundant in anthers but not in the pistil or leaf tissues. Among these miRNAs,

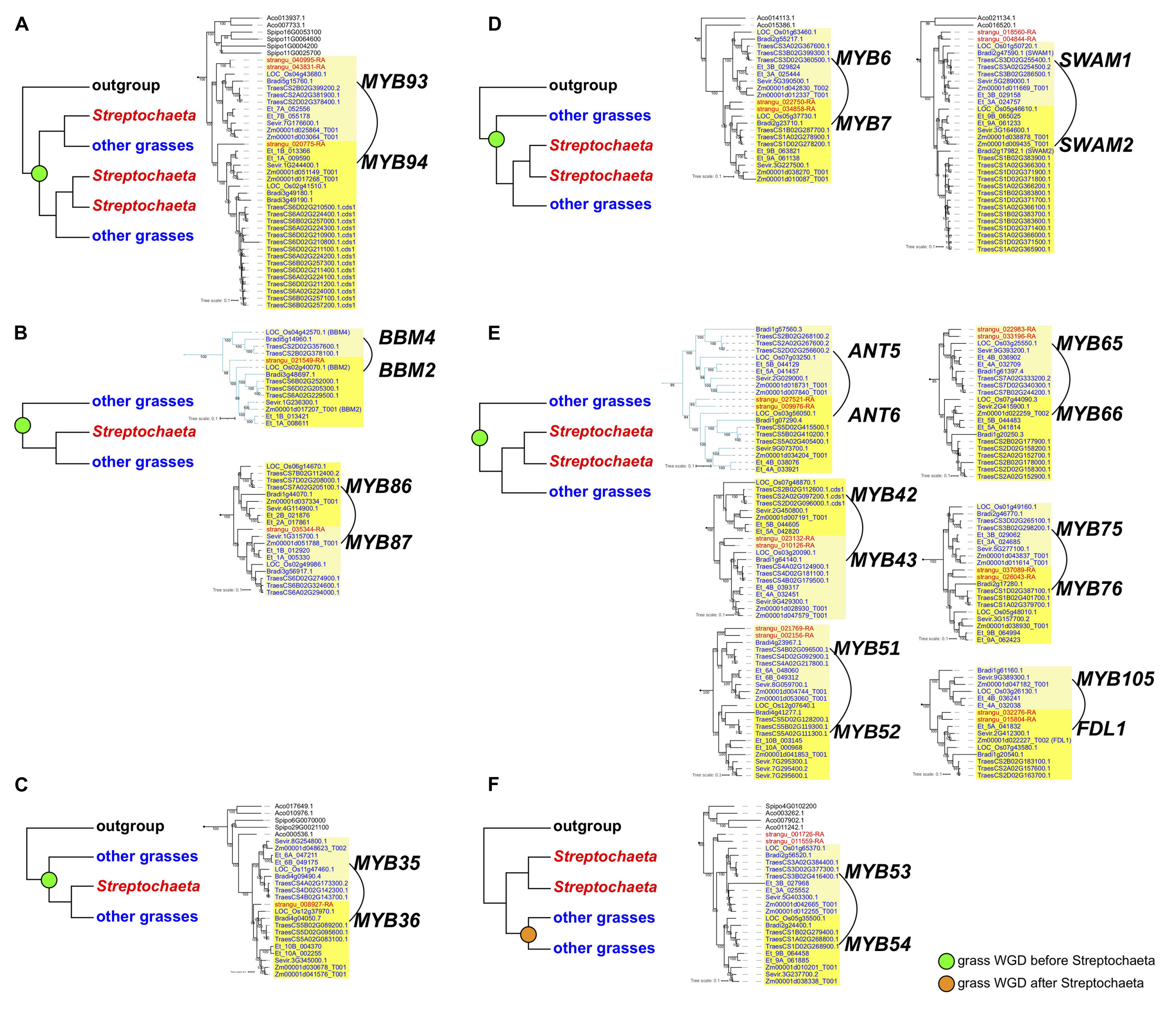

FIGURE 5 | Tree topologies of paired syntenic subclades that support grass whole genome duplication (WGD) before or after the divergence of *Streptochaeta*. **(A–E)** Grass WGD before the divergence of *Streptochaeta*. Tree topologies: **(A)** [O,(S1,G1),((S2,S3),G2)]. **(B)** [G1,(S2,G2)]. **(C)** [O,(G1,(S2,G2))]. **(D)** [O,(G1,((S1,S2),G2))]. **(E)** [G1,((S1,S2),G2)]. **(F)** Grass WGD after the divergence of *Streptochaeta* with tree pattern of [O,(S1,S2),(G1,G2)].

we found one copy each of miR2118 and miR2275, miRNAs known to function in the biogenesis of reproductive phasiRNAs (Johnson et al., 2009; Zhai et al., 2015). Among known miRNA families expressed in anthers, only 25.4% of families overlapped between Streptochaeta and three other monocots. The large number of miRNA families detected exclusively in anthers of asparagus (29.9%) and rice (17.9%) perhaps explains the small overlap between species.

We generated parallel analysis of RNA ends (PARE) libraries to identify and validate the cleavage of miRNA-target pairs in anther, pistil and leaf of *Streptochaeta* (**Supplementary Tables 11, 12**). Overall, we validated 58, 55, and 66 gene targets in anther, pistil and leaf, respectively. Half of these targets were detected in all tissues (51.9%), while 7 (8.6%), 4 (4.9%), and 14 (17.3%) were validated exclusively in anther, pistil, and leaf tissues, respectively; the remaining targets were found in combinations of two tissues. Among the validated targets, we found targets for three novel miRNAs, supporting their annotation. As an example, 184 reads validated the cleavage site of one novel miRNA target gene (strangu_031733), which is homologous to the *GPX6* gene (At4g11600), known to function in the protection of cells from oxidative damage in Arabidopsis (Rodriguez Milla et al., 2003). Among targets of known miRNAs, we validated the cleavage site of six and four genes encoding members of AP2 and MYB transcription factor families, respectively (**Supplementary Figures 2, 3**). miR172 triggered the cleavage of *AP2* genes in all tissues, consistent with the well-described function of this miRNA (Aukerman and Sakai, 2003; Lauter et al., 2005;

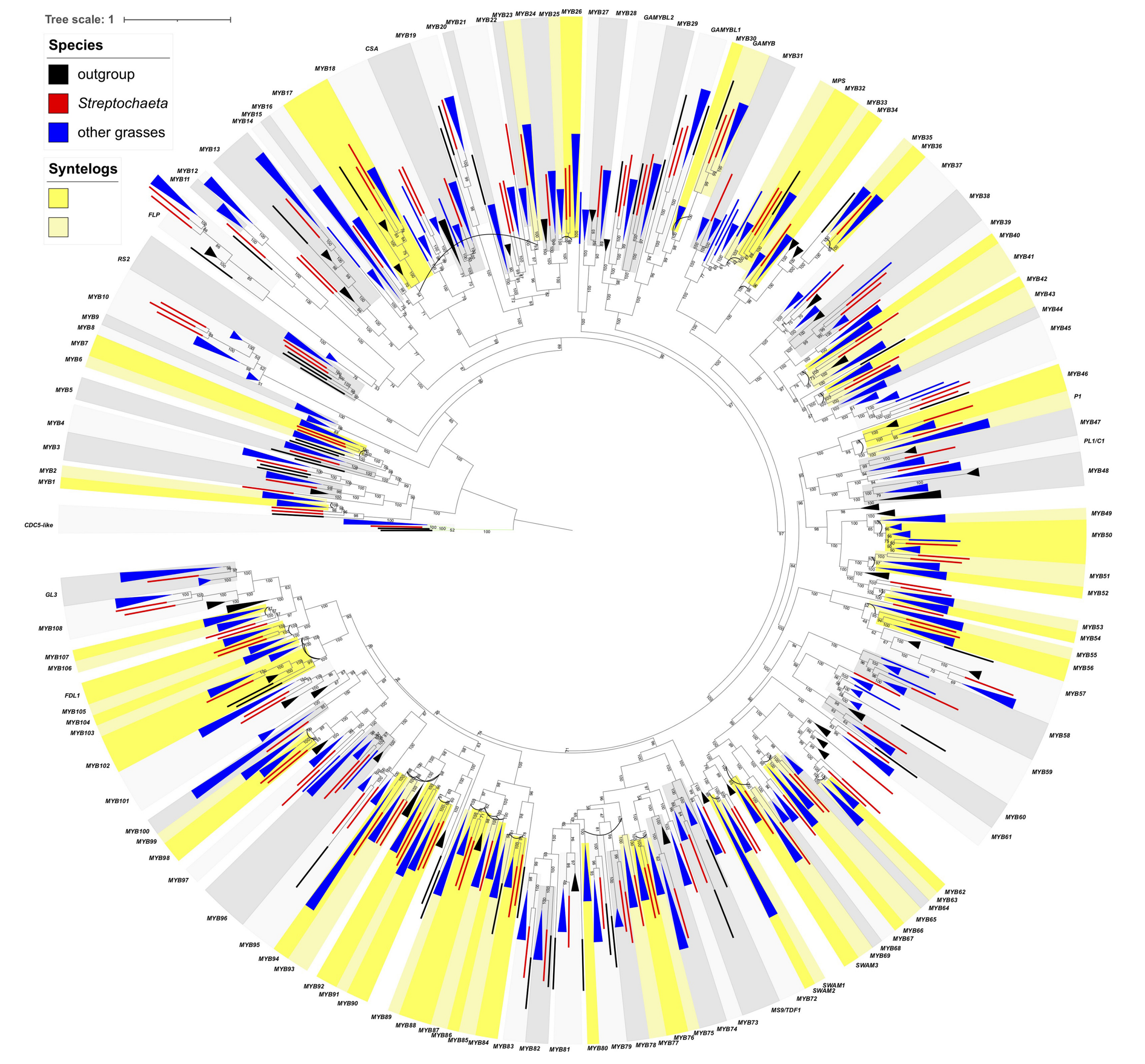

FIGURE 6 | Maximum likelihood tree of R2R3 genes. Numbers on branches indicate maximum likelihood bootstrap values. A single gene is denoted by a rectangle, and collapsed branches are denoted in triangles. Each subclade is shaded in two gray colors and named either by known genes within the subclade or subfamily name with a number. Subclades with syntenic genes in *Brachypodium*, *Oryza,* or *Setaria* are shaded in two colors of yellow, and syntenic pairs are connected by an arc. Outgroup, *Streptochaeta* and other grasses are shown in black, red, and blue colors.

Chuck et al., 2007, 2008). Also, miR159 triggered the cleavage of transcripts of four *MYB* genes homologous to rice *GAMYB* genes, in leaf and pistil tissues but not in anther.

Expression of PhasiRNAs Is Not Limited to Male Reproductive Tissues

phasiRNAs from the same sRNA libraries were annotated and the abundance of these loci compared to that in asparagus, maize, and rice. Overall, we detected a total of 89 phasiRNA loci (called *PHAS* loci) including 71 21-*PHAS* and 18 24-*PHAS* loci (**Supplementary Table 8**). We made three observations of note: First, we observed a switch in the ratio of 21-*PHAS* to 24-*PHAS* locus number comparing asparagus (<1), a member of Asparagaceae, to grass species (>1; Poaceae). Second, among Poaceae species, the number of genomic *PHAS* loci was lower in *Streptochaeta* than in both maize and rice. Third, several *PHAS* loci were also expressed in the pistil and leaves. Overall, 23 (32%) 21-*PHAS* loci and 11 (61%) 24-*PHAS* loci were expressed in the pistil with a median abundance of 32.9 and 12.3%, respectively,

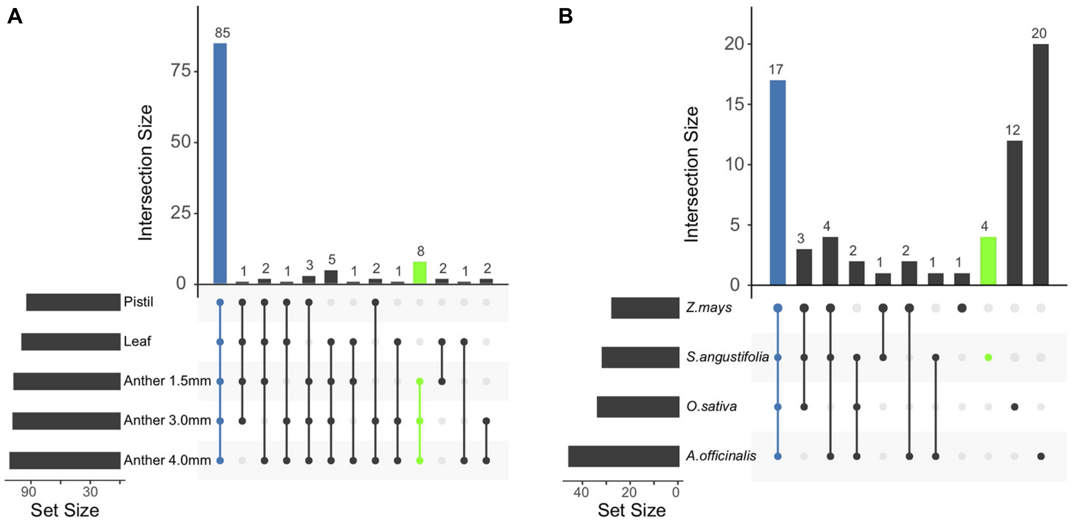

FIGURE 7 | Overlap of miRNA loci annotated in *Streptochaeta* tissues **(A)** and miRNA families annotated in *Streptochaeta* anthers compared to three other monocots **(B).**

compared to phasiRNAs detected in anther tissue. Similarly, 22 (31%) 21-*PHAS* loci and 10 (56%) 24-*PHAS* loci were detected in leaf tissue with a median abundance of 53.3 and 13.2%, respectively, compared to phasiRNAs detected in anthers. The expression of 24-nt phasiRNAs in vegetative tissues is unusual.

DISCUSSION

Genome Assembly and Annotation

The *Streptochaeta* genome presented here provides a resource for comparative genomics, genetics, and phylogenetics of the grass family. It represents the subfamily Anomochlooideae, which is sister to all other grasses and thus is equally phylogenetically distant to the better-known species rice, Brachypodium, sorghum, and maize (Clark et al., 1995; Grass Phylogeny Working Group [GPWG], 2001; Saarela et al., 2018). The genome assembly captures nearly all of the predicted gene space at high contiguity (complete BUSCOs 91.8%, liliopsida_odb10 profile, $n = 3278$), with the genome size matching predictions based on flow cytometry. The genome-wide LTR Assembly Index (LAI) for measuring the completeness of intact LTR elements, was 9.02, classifying the current genome as "draft" in quality, and is on par with other assemblies using similar sequencing technology [Apple (v1.0) (Velasco et al., 2010), Cacao (v1.0) (Argout et al., 2011)].

Our comprehensive annotation strategy identified a high proportion of genes specific to the genus *Streptochaeta*, also known as orphan genes (3,742). Many previous studies have indicated that orphan genes may comprise 3–10% of the total genes in plants and can, in certain species, range up to 30% of the total (Arendsee et al., 2014). Overall the average gene length (3,956 bp), average mRNA length (3,931 bp) and average CDS length (1,060 bp) are similar to other grass species queried in Ensembl (Howe et al., 2021).

Complex Evolutionary History of *Streptochaeta* May Contribute to Its Unique Characteristics

Our highly contiguous assembly in genic regions combined with gene model and functional annotations allowed: (1) evaluation of patterns of orthology between genes in *Streptochaeta* and BOP/PACMAD grasses to clarify the timing of the ρ WGD; (2) an investigation of gene families known to play a role in floral development that have potential relevance to the origin of the grass spikelet. Previous phylogenetic work based on transcriptomes (McKain et al., 2016) or individual gene tree analyses (Preston and Kellogg, 2006; Whipple et al., 2007; Christensen and Malcomber, 2012; McKain et al., 2016) suggested that *Streptochaeta* shared the same WGD (ρ) as the rest of the grasses but that it might also have its own duplication. Among the large sample (200) of clades in the transcriptome gene trees from McKain et al. (2016), 44% of these showed topologies consistent with ρ before the divergence of *Streptochaeta* (e.g., topologies shown in **Figures 1Ai,ii,iv**), with 39% being ambiguous (**Figures 1Aiii,Bii**). Fewer than 20% of the clades identified by McKain et al. (2016) had topologies consistent with the ρ duplication occurring after the divergence of *Streptochaeta* (**Figure 1Bi**). Additionally, *Streptochaeta* contigs show good collinearity with the rice genome, a finding that is consistent with ρ preceding the divergence of *Streptochaeta*. Mapping the *Streptochaeta* contigs against themselves also hints at another *Streptochaeta*-specific duplication, although the timing of this duplication cannot be inferred purely from the dot plot. Analyses of individual clades within large gene families (see below) support the same conclusion.

Analyzing the *AP2-like* and *MYB* subclades through the lens of grass WGD events, we found 12 and one cases supporting ρ before and after the divergence of *Streptochaeta,* thus confirming previous transcriptomic data (Preston and Kellogg, 2006;

Whipple et al., 2007; Christensen and Malcomber, 2012; McKain et al., 2016). We also found that *Streptochaeta* often lost one copy of the syntenic paralogs, not only in MADS-box genes (Preston and Kellogg, 2006; Christensen and Malcomber, 2012) but also in *AP2-like* and *R2R3 MYB* families. In addition, two *Streptochaeta* sequences are often sister to a grass clade (**Figure 5** and **Supplementary Table 7**), indicating that additional complexity in the evolution of Anomochlooideae may remain to be uncovered, although sequences of additional representatives of *S. angustifolia* as well as of the other three species in the subfamily will ultimately be needed.

Genome structure and phylogenetic trees of *Streptochaeta* genes and their orthologs support the "loss model" shown in **Figure 1Biv**, in which many of the genes known to control the structure of the grass spikelet were found in an ancestor of both *Streptochaeta* and the spikelet clade, but have then been lost in *Streptochaeta*. This provides circumstantial evidence that the common ancestor of all grasses – including *Streptochaeta* (and *Anomochloa*) – might have borne its flowers in spikelets, and the truly peculiar "spikelet equivalents" of Anomochlooideae are indeed highly modified.

Many transcription factor families are known to regulate spikelet development in the grasses (Hirano et al., 2014; Whipple, 2017). Of these, APETALA2 (AP2)-like proteins control meristem identity and floral morphology, including the number of florets per spikelet (Chuck et al., 1998; Lee and An, 2012; Zhou et al., 2012; Debernardi et al., 2020). Several R2R3 MYB proteins are also known to function in floral organ development, especially in anthers (Zhu et al., 2008; Aya et al., 2009; Zhang et al., 2010; Schmidt et al., 2013). We explored patterns of duplication and loss in these gene families between the stem node origin of the grasses and the origin of the spikelet clade, i.e., before and after the divergence of Streptochaeta.

Previous studies have focused on the evolution of MADS-box genes in shaping grass spikelet development. For example, the A-class gene in flower development *FRUITFULL* (*FUL*) duplicated at the base of Poaceae before the divergence of *Streptochaeta*, but *FUL1*/*VRN1* in *Streptochaeta* was subsequently lost (Preston and Kellogg, 2006). Similarly, paralogous *LEAFY HULL STERILE1* (*LHS1*) and *Oryza sativa MADS5* are duplicated at the base of Poaceae, but *Streptochaeta* has only one gene sister to the *LHS1* clade (Christensen and Malcomber, 2012). However, in another study on the B-class MADS-box gene *PISTILLATA* (*PI*), *Streptochaeta* has orthologs in both the *PI1* and *PI2* clades (Whipple et al., 2007).

Here, we focused on *AP2*-like and *R2R3 MYB* transcription factor families, both of which include members regulating inflorescence and spikelet development. The *euAP2* lineage of the *AP2*-like genes determines the transition from spikelet meristem to floral meristem (Hirano et al., 2014). In the maize mutant *indeterminate spikelet1* (*ids1*), extra florets are formed within the spikelets in both male and female flowers (Chuck et al., 1998). The double mutant of *ids1* and its syntenic paralog *sister of indeterminate spikelet1* (*sid1*) produce repetitive glumes (Chuck et al., 2008). Consistently, the rice mutants of *SUPERNUMERARY BRACT* (*SNB*), which is an ortholog of *SID1*, also exhibit multiple rudimentary glumes, due to the delay of transition from spikelet meristem to floral meristem. Such mutant phenotypes are somewhat analogous to the *Streptochaeta* "spikelet equivalents," which possess 11 or 12 bracts. *In situ* hybridization studies on *FUL* and *LHS1* showed that the outer bracts 1–5 resemble the expression pattern of glumes in other grass spikelets, while inner bracts 6–8 resemble the expression pattern of lemma and palea (Preston et al., 2009). Our phylogenetic analysis suggests that the ortholog of *IDS1* in *Streptochaeta* is lost (**Figure 4** and **Supplementary Figure 2**). Instead, *Streptochaeta* has two sequences orthologous to *SID1*/*SNB*, and these two sequences are successively sister to each other with a tree pattern of [G1,(S1,(S2,G2)] in *IDS1*/*Q-SID1*/*SNB* subclade pairs, leaving the evolutionary history of *Streptochaeta* ambiguous (**Figure 4**, **Supplementary Figure 2**, and **Supplementary Table 7**). Both *IDS1* and *SID1* are targets of miRNA172 in maize (Chuck et al., 2007, 2008). Our PARE analyses did validate the cleavage of all six *Streptochaeta euAP2* by miRNA172 (**Supplementary Table 12**), demonstrating that the miRNA172 post-transcriptional regulation of *euAP2* is functional in *Streptochaeta*. Detailed spatial gene expression analysis may further reveal whether and how these *euAP2* genes contribute to floral structure in *Streptochaeta*.

BABY BOOM genes *(BBMs)* belong to the euANT lineage of the *AP2*-like genes, and are well known for their function in induction of somatic embryogenesis (Boutilier et al., 2002) and application for *in vitro* tissue culture (Lowe et al., 2016). Ectopic expression of *BBM* in *Arabidopsis* and *Brassica* results in pleiotropic defects in plant development including changes in floral morphology (Boutilier et al., 2002). The grasses have four annotated *BBMs*, although it is not known whether other *ANT* members share similar functions. *BBM4* and *BBM2* subclades appeared to be duplicated paralog pairs due to the grass WGD. Similar to the cases in previous studies (Preston and Kellogg, 2006; Christensen and Malcomber, 2012), *Streptochaeta* has apparently lost its *BBM4* copy and contains one copy in the *BBM2* subclade (**Figures 4**, **5** and **Supplementary Figure 2**).

R2R3 MYB is a large transcription factor family, of which some members are crucial for anther development. The rice *carbon starved anther* (*csa*) mutants show decreased sugar content in floral organs including anthers, resulting in a male sterile phenotype (Zhang et al., 2010). *DEFECTIVE in TAPETAL DEVELOPMENT and FUNCTION1* (*TDF1*) is required for tapetum programmed cell death (Zhu et al., 2008; Cai et al., 2015). GAMYB positively regulates GA signaling by directly binding to the promoter of GA-responsive genes in both *Arabidopsis* and grasses (Tsuji et al., 2006; Aya et al., 2009; Alonso-Peral et al., 2010). *OsGAMYB* is highly expressed in stamen primordia, tapetum cells of the anther and aleurone cells, and its expression is regulated by miR159. Non-functional mutants of *OsGAMYB* are defective in tapetum development and are male sterile (Kaneko et al., 2004; Tsuji et al., 2006). We found conserved miRNA159 binding sites in *GAMYBs* and its closely related subclades, including *MYB27*, *MYB28*, *GAMYBL2*, *MYB29*, *GAMYBL1*, *MYB30,* and *GAMYB* (**Figure 4**). Our PARE analyses also validated the cleavage of *Streptochaeta GAMYB* and *GAMYBL1* in leaf and pistil tissues but not in anthers, suggesting the expression of *Streptochaeta GAMYB* and *GAMYBL1* may be suppressed by miR159 in tissues other than anthers, at least at the developmental stages we investigated (**Supplementary Table 12**). *Streptochaeta* has two sequences in each of the *GAMYBL2*, *MYB29*, *GAMYBL1* and *GAMYB* clades, either with a tree topology of [O,(S1,S2),G] in *GAMYBL2*, *MYB29,* and *GAMYBL1,* or a tree topology of [O,(S1,(S2,G)] in *GAMYB* (**Figures 4**, **6** and **Supplementary Table 7**). This again indicates that *Streptochaeta* has a complex duplication history.

A Survey of Small RNAs in the *Streptochaeta* Genome

sRNAs are important transcriptional and post-transcriptional regulators that play a role in plant development, reproduction, stress tolerance, etc. Identification of the complement of these molecules in *Streptochaeta* can inform our understanding of distinguishing features of grass and monocot genomes. miRNAs are major regulators of mRNA levels, active in pathways important to plant developmental transitions, biotic and abiotic stresses, and others. miRNAs generally act as post-transcriptional regulators by homology-dependent cleavage of target gene transcripts, when loaded to the RNA-induced silencing complex (RISC). Plant genomes encode a variety of sRNA types that can act in a transcriptional or post-transcriptional regulation mode. In this paper, we focused on miRNA and phasiRNA. The list of miRNA annotated in this study is likely incomplete because the *Streptochaeta* sRNA-seq data were limited to anther, pistil and leaf tissues, and would miss miRNAs expressed specifically in other tissues/cell types or at growth conditions not sampled. Thus, miRNAs missed in our data may well be encoded in the *Streptochaeta* genome. That being said, our miRNA characterization provides a starting point with which to describe *Streptochaeta* miRNAs, and our sequencing depth and tissue diversity was likely sufficient to identify many if not the majority of miRNAs encoded in the genome.

Phased short interfering RNAs (phasiRNAs) are 21-nt or 24-nt sRNAs generated from the recursive cleavage of a double-stranded RNA from a well-defined terminus; these transcripts define their precursor *PHAS* loci (Axtell and Meyers, 2018). Reproductive phasiRNAs are a subset abundant in anthers and in some cases essential to male fertility. Genomes of grass species are particularly rich in reproductive *PHAS* loci (Patel et al., 2021), expressed in anthers but not in female reproductive tissues or vegetative tissues. Previous species studies identified hundreds of *PHAS* loci in anthers of maize (Zhai et al., 2015) to thousands of *PHAS* loci in rice (Fei et al., 2016), barley (Bélanger et al., 2020), and bread wheat (Bélanger et al., 2020; Zhang et al., 2020). Additionally, work in maize (Teng et al., 2020) and rice (Fan et al., 2016) showed that 21-nt and 24-nt phasiRNAs are essential to ensure proper development of meiocytes and to guarantee male fertility under normal growth conditions. However, *Streptochaeta* has a different internal anatomy than the rest of the grasses. Specifically, anthers in *Streptochaeta* are missing the "middle layer" between the endothecium and the tapetum (Sajo et al., 2009, 2012) such that the microsporangium has only three cell layers.

Given that most of our data (>100 M reads) were collected from anthers, we have good resolution for annotation of phasiRNAs in this tissue. We characterized their absence/presence in the three-layer anthers of *Streptochaeta*. We annotated tens of *PHAS* loci in *Streptochaeta* showing that anthers express phasiRNAs even in the absence of the middle layer. Likewise, in maize, Zhai et al. (2015) showed that the miRNA and phasiRNA precursors are dependent on the epidermis, endothecium, and tapetum, and the phasiRNAs accumulate in the tapetum and meiocytes, so the middle layer is apparently not involved. We observed a shift in the ratio of 21-*PHAS* to 24-*PHAS* loci from asparagus (<1), an Asparagaceae, to grass species (>1), although the implications of this shift are as yet unclear.

We also observed that several 21-nt and 24-nt phasiRNAs accumulate in either pistil or leaf tissues, inconsistent with prior results. A small number of 21-nt *PHAS* loci are likely trans-acting-siRNA-generating (*TAS*) loci, important in vegetative tissues, but typically there are only a few *TAS* loci per genome (Xia et al., 2017), not the 20 loci that we observed. Additionally, we found no previous reports of 24-nt phasiRNAs accumulating in vegetative tissues or female reproductive tissues.

Utility of *Streptochaeta* for Understanding Grass Evolution and Genetics

The four species of Anomochlooideae contribute to understanding the evolution of the grasses and the many traits that make them unique. We have highlighted the unusual floral and inflorescence morphology of *Streptochaeta* and have compared it to grass spikelets, but *Streptochaeta* can also illuminate the evolution and genetic basis of other important traits. It is common to compare traits between members of the BOP clade (e.g., *Oryza*, *Brachypodium*, or *Triticum*) and the PACMAD clade (e.g., *Zea*, *Sorghum*, *Panicum*, *Eragrostis*), but, because these comparisons involve two sister clades, it is impossible to determine whether the BOP or the PACMAD clade character state is ancestral. *Streptochaeta* functions as an outgroup in such comparisons and can help establish the direction of change. Here, we highlight just a few of the traits whose analysis may be helped in future studies by reference to *Streptochaeta* and its genome sequence.

Drought Intolerance, Shade Tolerance

The grasses, including not only Anomochlooideae, but also Pharoideae and Puelioideae, the three subfamilies that are successive sister groups of the rest of the family, appear to have originated in environments with low light and high humidity (Edwards and Smith, 2010; Gallaher et al., 2019). The shift from shady, moist habitats to open, dry habitats where most grass species are now found promises insights into photosynthesis and water use efficiency, among other physiological traits.

Streptochaeta, like other forest grasses, has broad, spreading leaf blades and a pseudopetiole that results in higher leaf angle and increased light interception (Gallaher et al., 2019). Leaf angle is an important agronomic trait, with selection during modern breeding often favoring reduced leaf angle to maximize plant density and yield (Liu et al., 2019; Mantilla-Perez et al., 2020). A close examination of *Streptochaeta* may provide insight into how leaf angle is controlled in diverse grasses. Leaf width in maize is controlled particularly by the *WOX3*-like homeodomain proteins *NARROWSHEATH1* (*NS1*) and *NS2*, which function in cells at the margins of leaves (Scanlon et al., 1996; Conklin et al., 2020). Duplication patterns and expression of *NS1* and *NS2* genes in the *Streptochaeta* genome could test whether the models developed for maize were present in the earliest of grasses.

Leaf Anatomy

The grass outgroup *Joinvillea* develops colorless cells in the mesophyll (Leandro et al., 2018). These appear to form from the same ground tissue that is responsible for the cavity-like "fusoid" cells in Anomochlooideae, Pharoideae, and Puelioideae as well as the bambusoid grasses. These cells, which appear to be a shared derived character for the grasses, form from the collapse of mesophyll cells and may play a role in the synthesis and storage of starch granules early in plant development (Leandro et al., 2018). While the genetic basis of leaf anatomy is, at the moment, poorly understood, *Streptochaeta* will be a useful system for understanding the development of fusoid cells in early diverging and other grasses.

Grass leaves also contain silica bodies in the epidermis; the vacuoles of these cells are filled with amorphous silica (SiO_2). In *Streptochaeta* the silica bodies are a distinctive shape, being elongated transverse to the long axis of the blade (Judziewicz and Soderstrom, 1989). The genetic basis of silica deposition has been studied in rice (Yu et al., 2020) and the availability of the *Streptochaeta* genome now permits examination of the evolution of these genes in the grasses.

Anther and Pollen Development

Streptochaeta differs from most other grasses (and indeed some Poales as well) in details of its anthers and pollen development, and the current genome provides tools for comparative analyses. The sRNAs described above are produced in the epidermis, endothecium and tapetum of most grasses and we presume they are also produced in those tissues in *Streptochaeta*. In all grasses except Anomochlooideae and Pharoideae, the microsporangium has four concentric layers of cells – the epidermis, the endothecium, the middle layer, and the tapetum – which surround the archesporial cells (Walbot and Egger, 2016). Cells in the middle layer and the tapetum are sisters, derived from division of a secondary parietal cell. The inner walls of the endothecial cells also mature to become fibrous (Artschwager and McGuire, 1949; Furness and Rudall, 1998). In *Streptochaeta* and *Pharus*, however, the middle layer is absent (Sajo et al., 2007, 2009, 2012) and the endothecial cells lack fibrous thickenings. It is tempting to speculate that the middle layer may have a role in coordinating maturation of the endothecium. Lack of the middle layer is apparently derived within *Streptochaeta* and *Pharus*. In known mutants of maize and rice, loss of the middle layer leads to male sterility (Walbot and Egger, 2016) so the functional implications of its absence in *Streptochaeta* are unclear.

Development of microsporangium layers may also be related to the position of microspores inside the locule. In most grasses, the microspores and mature pollen grains form a single layer adjacent to the tapetum, with the pore of the pollen grain facing the tapetum, unlike many non-grasses in which the microsporocytes fill the locule and have a haphazard arrangement. The condition in *Streptochaeta* is unclear, with contradictory reports in the literature (Kirpes et al., 1996; Sajo et al., 2009, 2012).

The exine, or outer layer, of grass pollen is distinct from that of its close relatives due to the presence of channels that pass through the exine. While controls of this particular aspect of the pollen wall are unknown in the grasses, we find that *Streptochaeta* and its grass sisters have several GAMYB genes, which are known to be involved in exine formation in rice (Aya et al., 2009) and to have played a role more broadly in reproductive processes, including microspore development in early vascular plants (Aya et al., 2011).

Chromosome Number in the Early Grasses

Estimates of the chromosome number and karyotype in the common ancestor of grasses have reached different conclusions (e.g., Salse et al., 2008; Murat et al., 2010; Wang et al., 2016), in part because of limited taxon sampling particularly for early diverging lineages, which heavily affect optimization of any characters. Genomes of *Streptochaeta* and other early diverging grasses will be useful for resolving this open question, but will require pseudomolecule-quality assemblies. Two other species of *Streptochaeta* have been reported to have n =11 chromosomes (Valencia, 1962; Pohl and Davidse, 1971; Hunziker et al., 1982), well below the number reported for the sister species *Anomochloa marantoidea*, $n = 18$ (Judziewicz and Soderstrom, 1989). The outgroups *Joinvillea plicata* and *Ecdeiocolea monostachya* have $n = 18$ (Newell, 1969) and $n = 19$ (Hanson et al., 2005), respectively. However, without high quality genomes and good cytogenetic data for these species, the ancestral chromosome number and structure of the genomes of ancestral grasses remains a matter of speculation.

Finally, these are but a few of the opportunities for understanding trait evolution in the grasses based on investigation of *Streptochaeta*, with additional insights possible in, for example, the study of embryo development, caryopsis modifications, endosperm/starch evolution and branching/tillering. We have demonstrated that genomes of targeted, non-model species, particularly those that are sister to large, better-studied groups, can provide out-sized insight about the nature of evolutionary transitions and should be an increased focus now that genome assembly is a broadly accessible component of the biologist's toolkit.

DATA AVAILABILITY STATEMENT

The sRNA-seq data were reported in a previous study (Patel et al., 2021). Also, one library of RNA-Seq (SRR3233339) used for annotation was previously published (Givnish et al., 2010). Otherwise, all data utilized in this study are original. The complete set of raw WGS, RNA-seq, sRNA-seq, PARE-seq reads and the genome assembly were deposited in NCBI under BioProject ID PRJNA343128. The final genome assembly, annotation and Bionano optical maps are also available in CyVerse data commons at: https://datacommons.cyverse.org/browse/iplant/home/aseetharam/Streptochaeta_v1_publication_release_2021-05-13. The scripts and commands used for generating assembly, annotations, small RNA analyses and phylogenetic analyses are documented in the GitHub repository: https://github.com/HuffordLab/streptochaeta.

AUTHOR CONTRIBUTIONS

MH, AS, EK, and LC designed the project. LC and EK provided the plant material. MH and AS generated the sequence data and assembled the genome. SB and BM analyzed the data on small RNAs. YY analyzed the AP2 and MYB sequence data. All authors drafted and edited the manuscript, and produced figures and tables.

FUNDING

YY was supported by NSF grant IOS-1938086 to EK and by an Enterprise-Rent-a-Car Foundation award through the Donald Danforth Plant Science Center, also to EK. SB was supported by USDA | National Institute of Food and Agriculture "BTT EAGER" award no. 2018–09058 to BM, as well as resources from the Donald Danforth Plant Science Center and the University of Missouri–Columbia. AS was supported by NSF grant IOS-1822330 to MH. This work used (1) Extreme Science and Engineering Discovery Environment (XSEDE) (National Science Foundation Grant No. ACI-1548562) via Blacklight HPC environment allocation TG-MCB140103 and (2) HPC equipment at Iowa State University, some of which has been purchased through funding provided by NSF under MRI grant number 1726447.

ACKNOWLEDGMENTS

We thank Sandra Mathioni for construction of the RNA-seq and PARE libraries. We also thank Philip Blood for his assistance with MaSuRCA optimization, which was made possible through the XSEDE Extended Collaborative Support Service (ECSS) program.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2021.710383/full#supplementary-material

Supplementary Figure 1 | Dot plots depicting whole genome alignments of *Streptochaeta* scaffolds with rice chromosomes. Dots aligned diagonally shows conserved synteny between these genomes.

Supplementary Figure 2 | Maximum likelihood tree of *AP2*-like genes with gene names. Bootstrap values are shown on the branches. Each subclade is shaded in two gray colors and named either by known genes within the subclade or subfamily name with a number. Subclades with syntenic genes in *Brachypodium*, *Oryza,* or *Setaria* are shaded in two colors of yellow, and syntenic pairs are connected by an arc. Predicted and experimentally validated miR172 binding sites are denoted by red and green stars, respectively.

Supplementary Figure 3 | Maximum likelihood tree of *R2R3* genes with gene names. Bootstrap values are shown on the branches. Each subclade is shaded in two gray colors and named either by known genes within the subclade or subfamily name with a number. Subclades with syntenic genes in *Brachypodium*, *Oryza,* or *Setaria* are shaded in two colors of yellow, and syntenic pairs are connected by an arc. Predicted and experimental validated miR159 binding sites are denoted by red and green stars, respectively.

REFERENCES

Alonso-Peral, M. M., Li, J., Li, Y., Allen, R. S., Schnippenkoetter, W., Ohms, S., et al. (2010). The microRNA159-regulated GAMYB-like genes inhibit growth and promote programmed cell death in Arabidopsis. *Plant Physiol.* 154, 757–771. doi: 10.1104/pp.110.160630

Arendsee, Z., Li, J., Singh, U., Seetharam, A., Dorman, K., and Wurtele, E. S. (2019). phylostratr: a framework for phylostratigraphy. *Bioinformatics* 35, 3617–3627. doi: 10.1093/bioinformatics/btz171

Arendsee, Z. W., Li, L., and Wurtele, E. S. (2014). Coming of age: orphan genes in plants. *Trends Plant Sci.* 19, 698–708. doi: 10.1016/j.tplants.2014.07.003

Argout, X., Salse, J., Aury, J.-M., Guiltinan, M. J., Droc, G., Gouzy, J., et al. (2011). The genome of *Theobroma cacao*. *Nat. Genet.* 43, 101–108. doi: 10.1038/ng.736

Artschwager, E., and McGuire, R. C. (1949). Cytology of reproduction in *Sorghum vulgare*. *J. Agric. Res.* 78, 659–673.

Aukerman, M. J., and Sakai, H. (2003). Regulation of flowering time and floral organ identity by a MicroRNA and its APETALA2-like target genes. *Plant Cell* 15, 2730–2741. doi: 10.1105/tpc.016238

Axtell, M. J., and Meyers, B. C. (2018). Revisiting criteria for plant MicroRNA annotation in the era of big data. *Plant Cell* 30, 272–284. doi: 10.1105/tpc.17.00851

Aya, K., Hiwatashi, Y., Kojima, M., Sakakibara, H., Ueguchi-Tanaka, M., Hasebe, M., et al. (2011). The Gibberellin perception system evolved to regulate a pre-existing GAMYB-mediated system during land plant evolution. *Nat. Commun.* 2:544. doi: 10.1038/ncomms1552

Aya, K., Ueguchi-Tanaka, M., Kondo, M., Hamada, K., Yano, K., Nishimura, M., et al. (2009). Gibberellin modulates anther development in rice via the transcriptional regulation of GAMYB. *Plant Cell* 21, 1453–1472. doi: 10.1105/tpc.108.062935

Bartlett, M., Thompson, B., Brabazon, H., Del Gizzi, R., Zhang, T., and Whipple, C. (2016). Evolutionary dynamics of floral homeotic transcription factor protein-protein interactions. *Mol. Biol. Evol.* 33, 1486–1501. doi: 10.1093/molbev/msw031

Bélanger, S., Pokhrel, S., Czymmek, K., and Meyers, B. C. (2020). Premeiotic, 24-nucleotide reproductive PhasiRNAs are abundant in anthers of wheat and barley but not rice and maize. *Plant Physiol.* 184, 1407–1423. doi: 10.1101/2020.06.18.160440

Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D., and Pirovano, W. (2011). Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* 27, 578–579. doi: 10.1093/bioinformatics/btq683

Bonnet, E., He, Y., Billiau, K., and Van de Peer, Y. (2010). TAPIR, a web server for the prediction of plant microRNA targets, including target mimics. *Bioinformatics* 26, 1566–1568. doi: 10.1093/bioinformatics/btq233

Bouchenak-Khelladi, Y., Onstein, R. E., Xing, Y., Schwery, O., and Linder, H. P. (2015). On the complexity of triggering evolutionary radiations. *New Phytol.* 207, 313–326. doi: 10.1111/nph.13331

Boutilier, K., Offringa, R., Sharma, V. K., Kieft, H., Ouellet, T., Zhang, L., et al. (2002). Ectopic expression of BABY BOOM triggers a conversion from vegetative to embryonic growth. *Plant Cell* 14, 1737–1749. doi: 10.1105/tpc.001941

Cai, C.-F., Zhu, J., Lou, Y., Guo, Z.-L., Xiong, S.-X., Wang, K., et al. (2015). The functional analysis of OsTDF1 reveals a conserved genetic pathway for tapetal development between rice and Arabidopsis. *Sci. Bull. Fac. Agric. Kyushu Univ.* 60, 1073–1082. doi: 10.1007/s11434-015-0810-3

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., and Bealer, K. (2009). BLAST plus: architecture and applications. *BMC Bioinform.* 10:421. doi: 10.1186/1471-2105-10-421

Capella-Gutiérrez, S., Silla-Martínez, J. M., and Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25, 1972–1973. doi: 10.1093/bioinformatics/btp348

Christensen, A. R., and Malcomber, S. T. (2012). Duplication and diversification of the LEAFY HULL STERILE1 and *Oryza sativa* MADS5 SEPALLATA lineages in graminoid Poales. *Evodevo* 3:4. doi: 10.1186/2041-9139-3-4

Chuck, G., Meeley, R., and Hake, S. (2008). Floral meristem initiation and meristem cell fate are regulated by the maize AP2 genes *ids1* and *sid1*. *Development* 135, 3013–3019. doi: 10.1242/dev.024273

Chuck, G., Meeley, R., Irish, E., Sakai, H., and Hake, S. (2007). The maize *tasselseed4* microRNA controls sex determination and meristem cell fate by targeting *Tasselseed6/indeterminate spikelet1*. *Nat. Genet.* 39, 1517–1521. doi: 10.1038/ng.2007.20

Chuck, G., Meeley, R. B., and Hake, S. (1998). The control of maize spikelet meristem fate by the APETALA2-like gene *indeterminate spikelet1*. *Genes Dev.* 12, 1145–1154. doi: 10.1101/gad.12.8.1145

Clark, L. G., Zhang, W., and Wendel, J. F. (1995). A phylogeny of the grass family (Poaceae) based on *ndhF* sequence data. *Syst. Bot.* 20, 436–460. doi: 10.2307/2419803

Conklin, P. A., Johnston, R., Conlon, B. R., Shimizu, R., and Scanlon, M. J. (2020). Plant homeodomain proteins provide a mechanism for how leaves grow wide. *Development* 147:dev193623. doi: 10.1242/dev.193623

Conway, J. R., Lex, A., and Gehlenborg, N. (2017). UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics* 33, 2938–2940. doi: 10.1093/bioinformatics/btx364

Debernardi, J. M., Greenwood, J. R., Jean Finnegan, E., Jernstedt, J., and Dubcovsky, J. (2020). APETALA 2-like genes *AP2L2* and *Q* specify lemma identity and axillary floral meristem development in wheat. *Plant J.* 101, 171–187. doi: 10.1111/tpj.14528

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26, 2460–2461. doi: 10.1093/bioinformatics/btq461

Edwards, E. J., and Smith, S. A. (2010). Phylogenetic analyses reveal the shady history of C4 grasses. *Proc. Natl. Acad. Sci. U.S.A.* 107, 2532–2537. doi: 10.1073/pnas.0909672107

Fahlgren, N., and Carrington, J. C. (2010). miRNA target prediction in plants. *Methods Mol. Biol.* 592, 51–57. doi: 10.1007/978-1-60327-005-2_4

Fan, Y., Yang, J., Mathioni, S. M., Yu, J., Shen, J., Yang, X., et al. (2016). PMS1T, producing phased small-interfering RNAs, regulates photoperiod-sensitive male sterility in rice. *Proc. Natl. Acad. Sci. U.S.A.* 113, 15144–15149. doi: 10.1073/pnas.1619159114

Fei, Q., Yang, L., Liang, W., Zhang, D., and Meyers, B. C. (2016). Dynamic changes of small RNAs in rice spikelet development reveal specialized reproductive phasiRNA pathways. *J. Exp. Bot.* 67, 6037–6049. doi: 10.1093/jxb/erw361

Furness, C. A., and Rudall, P. J. (1998). The tapetum and systematics in monocotyledons. *Bot. Rev.* 64, 201–239. doi: 10.1007/BF02856565

Gallaher, T. J., Adams, D. C., Attigala, L., Burke, S. V., Craine, J. M., Duvall, M. R., et al. (2019). Leaf shape and size track habitat transitions across forest-grassland boundaries in the grass family (Poaceae). *Evolution* 73, 927–946. doi: 10.1111/evo.13722

Gibson, D. J. (2009). *Grasses and Grassland Ecology*. Oxford, UK: Oxford University Press.

Givnish, T. J., Ames, M., McNeal, J. R., McKain, M. R., Roxanne Steele, P., dePamphilis, C. W., et al. (2010). Assembling the tree of the monocotyledons: plastome sequence phylogeny and evolution of Poales. *Ann. Mo. Bot. Gard.* 97, 584–616. doi: 10.3417/2010023

Grass Phylogeny Working Group [GPWG] (2001). Phylogeny and subfamilial classification of the grasses (Poaceae). *Ann. Mo. Bot. Gard.* 88, 373–457. doi: 10.2307/3298585

Grass Phylogeny Working Group II [GPWG II] (2012). New grass phylogeny resolves deep evolutionary relationships and discovers C4 origins. *New Phytol.* 193, 304–312. doi: 10.1111/j.1469-8137.2011.03972.x

Hanson, L., Boyd, A., Johnson, M. A. T., and Bennett, M. D. (2005). First nuclear DNA C-values for 18 eudicot families. *Ann. Bot.* 96, 1315–1320. doi: 10.1093/aob/mci283

Haug-Baltzell, A., Stephens, S. A., Davey, S., Scheidegger, C. E., and Lyons, E. (2017). SynMap2 and SynMap3D: web-based whole-genome synteny browsers. *Bioinformatics* 33, 2197–2198. doi: 10.1093/bioinformatics/btx144

Hirano, H.-Y., Tanaka, W., and Toriba, T. (2014). "Grass flower development," in *Flower Development: Methods and Protocols*, eds J. L. Riechmann and F. Wellmer (New York, NY: Springer New York), 57–84. doi: 10.1007/978-1-4614-9408-9_3

Hoff, K. J., Lomsadze, A., Borodovsky, M., and Stanke, M. (2019). "Whole-Genome annotation with BRAKER," in *Gene Prediction: Methods and Protocols*, ed. M. Kollmar (New York, NY: Springer New York), 65–95. doi: 10.1007/978-1-4939-9173-0_5

Howe, K. L., Achuthan, P., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M. R., et al. (2021). Ensembl 2021. *Nucleic Acids Res.* 49, D884–D891. doi: 10.1093/nar/gkaa942

Huang, Y., Zhao, S., Fu, Y., Sun, H., Ma, X., Tan, L., et al. (2018). Variation in the regulatory region of *FZP* causes increases in secondary inflorescence branching and grain yield in rice domestication. *Plant J.* 96, 716–733. doi: 10.1111/tpj.14062

Hunziker, J. H., Wulff, A. F., and Soderstrom, T. R. (1982). Chromosome studies on the Bambusoideae (Gramineae). *Brittonia* 34:30. doi: 10.2307/2806397

International Brachypodium Initiative [IBI] (2010). Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature* 463, 763–768. doi: 10.1038/nature08747

Jiang, C.-K., and Rao, G.-Y. (2020). Insights into the diversification and evolution of R2R3-MYB transcription factors in plants. *Plant Physiol.* 183, 637–655. doi: 10.1104/pp.19.01082

Johnson, C., Kasprzewska, A., Tennessen, K., Fernandes, J., Nan, G.-L., Walbot, V., et al. (2009). Clusters and superclusters of phased small RNAs in the developing inflorescence of rice. *Genome Res.* 19, 1429–1440. doi: 10.1101/gr.089854.108

Johnson, N. R., Yeoh, J. M., Coruh, C., and Axtell, M. J. (2016). Improved placement of multi-mapping small RNAs. *G3* 6, 2103–2111. doi: 10.1534/g3.116.030452

Judziewicz, E. J., Clark, L. G., Londoño, X., and Stern, M. J. (1999). *American Bamboos*. Washington, DC: Smithsonian Books.

Judziewicz, E. J., and Soderstrom, T. R. (1989). Morphological, Anatomical, and Taxonomic Studies in *Anomochloa* and *Streptochaeta* (Poaceae: Bambusoideae). Smithson. *Contr. Bot.* 68, 1–52.

Kaneko, M., Inukai, Y., Ueguchi-Tanaka, M., Itoh, H., Izawa, T., Kobayashi, Y., et al. (2004). Loss-of-function mutations of the rice *GAMYB* gene impair alpha-amylase expression in aleurone and flower development. *Plant Cell* 16, 33–44. doi: 10.1105/tpc.017327

Katoh, K., and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780. doi: 10.1093/molbev/mst010

Kellogg, E. A. (2015). "Poaceae," in *The Families and Genera of Vascular Plants*, ed. K. Kubitzki (Berlin: Springer), 1–416.

Kellogg, E. A., Camara, P. E. A. S., Rudall, P. J., Ladd, P., Malcomber, S. T., Whipple, C. J., et al. (2013). Early inflorescence development in the grasses (Poaceae). *Front. Plant Sci.* 4:250. 10.3389/fpls.2013.00250

Kielbasa, S. M., Wan, R., Sato, K., Horton, P., and Frith, M. C. (2011). Adaptive seeds tame genomic sequence comparison. *Genome Res.* 21, 487–493. doi: 10.1101/gr.113985.110

Kim, S., Soltis, P. S., Wall, K., and Soltis, D. E. (2006). Phylogeny and domain evolution in the APETALA2-like gene family. *Mol. Biol. Evol.* 23, 107–120. doi: 10.1093/molbev/msj014

Kirpes, C. C., Clark, L. G., and Lersten, N. R. (1996). Systematic significance of pollen arrangement in microsporangia of Poaceae and Cyperaceae: review and observations on representative taxa. *Am. J. Bot.* 83, 1609–1622. doi: 10.1002/j.1537-2197.1996.tb12819.x

Kozomara, A., Birgaoanu, M., and Griffiths-Jones, S. (2019). miRBase: from microRNA sequences to function. *Nucleic Acids Res.* 47, D155–D162. doi: 10.1093/nar/gky1141

Kozomara, A., and Griffiths-Jones, S. (2014). miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.* 42, D68–D73. doi: 10.1093/nar/gkt1181

Laetsch, D. R., and Blaxter, M. L. (2017). BlobTools: interrogation of genome assemblies. *F1000Res* 6:1287. doi: 10.12688/f1000research.12232.1

Lauter, N., Kampani, A., Carlson, S., Goebel, M., and Moose, S. P. (2005). *microRNA172* down-regulates *glossy15* to promote vegetative phase change in maize. *Proc. Natl. Acad. Sci. U.S.A.* 102, 9412–9417. doi: 10.1073/pnas.0503927102

Leandro, T. D., Rodrigues, T. M., Clark, L. G., and Scatena, V. L. (2018). Fusoid cells in the grass family Poaceae (Poales): a developmental study reveals homologies and suggests new insights into their functional role in young leaves. *Ann. Bot.* 122, 833–848. doi: 10.1093/aob/mcy025

Lee, D.-Y., and An, G. (2012). Two AP2 family genes, *supernumerary bract (SNB)* and *Osindeterminate spikelet 1 (OsIDS1)*, synergistically control inflorescence architecture and floral meristem establishment in rice. *Plant J.* 69, 445–461. doi: 10.1111/j.1365-313X.2011.04804.x

Lehmann, C. E. R., Griffith, D. M., Simpson, K. J., Michael Anderson, T., Archibald, S., Beerling, D. J., et al. (2019). Functional diversification enabled grassy biomes to fill global climate space. *Biorxiv [Preprint]* 10.1101/583625

Letunic, I., and Bork, P. (2019). Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.* 47, W256–W259. doi: 10.1093/nar/gkz239

Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R., and Pfister, H. (2014). UpSet: visualization of intersecting sets. *IEEE Trans. Vis. Comput. Graph.* 20, 1983–1992. doi: 10.1109/TVCG.2014.2346248

Li, C., Lin, H., Chen, A., Lau, M., Jernstedt, J., and Dubcovsky, J. (2019). Wheat VRN1, FUL2 and FUL3 play critical and redundant roles in spikelet development and spike determinacy. *Development* 146:dev175398. 10.1242/dev.175398

Li, Y., Zhu, J., Wu, L., Shao, Y., Wu, Y., and Mao, C. (2019). Functional divergence of *PIN1* paralogous genes in rice. *Plant Cell Physiol.* 60, 2720–2732. doi: 10.1093/pcp/pcz159

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34, 3094–3100. doi: 10.1093/bioinformatics/bty191

Linder, H. P., Lehmann, C. E. R., Archibald, S., Osborne, C. P., and Richardson, D. M. (2018). Global grass (Poaceae) success underpinned by traits facilitating colonization, persistence and habitat transformation. *Biol. Rev. Camb. Philos. Soc.* 93, 1125–1144. doi: 10.1111/brv.12388

Liu, K., Cao, J., Yu, K., Liu, X., Gao, Y., Chen, Q., et al. (2019). Wheat *TaSPL8* modulates leaf angle through auxin and brassinosteroid signaling [OPEN]. *Plant Physiol.* 181, 179–194. doi: 10.1104/pp.19.00248

Lorenz, R., Bernhart, S. H., Höner Zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P. F., et al. (2011). ViennaRNA Package 2.0. *Algorithms Mol. Biol.* 6:26. doi: 10.1186/1748-7188-6-26

Lowe, K., Wu, E., Wang, N., Hoerster, G., Hastings, C., Cho, M.-J., et al. (2016). Morphogenic regulators BABY BOOM and WUSCHEL improve monocot transformation. *Plant Cell* 28, 1998–2015. doi: 10.1105/tpc.16.00124

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., et al. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* 1:18. doi: 10.1186/2047-217X-1-18

Lyons, E., and Freeling, M. (2008). How to usefully compare homologous plant genes and chromosomes as DNA sequences. *Plant J.* 53, 661–673. doi: 10.1111/j.1365-313X.2007.03326.x

Magallón, S. Gómez-Acevedo, S., Sánchez-Reyes, L. L., and Hernández-Hernández, T. (2015). A metacalibrated time-tree documents the early rise of the flowering plant phylogenetic diversity. *New Phytol.* 207, 437–453. doi: 10.1111/nph.13264

Mamidi, S., Healey, A., Huang, P., Grimwood, J., Jenkins, J., Barry, K., et al. (2020). A genome resource for green millet *Setaria viridis* enables discovery of agronomically valuable loci. *Nat. Biotechnol.* 38, 1203–1210. doi: 10.1038/s41587-020-0681-2

Mantilla-Perez, M. B., Bao, Y., Tang, L., Schnable, P. S., and Salas-Fernandez, M. G. (2020). Toward "Smart Canopy" sorghum: discovery of the genetic control of leaf angle across layers. *Plant Physiol.* 184, 1927–1940. doi: 10.1104/pp.20.00632

Marçais, G., and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 27, 764–770. doi: 10.1093/bioinformatics/btr011

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal* 17, 10–12. doi: 10.14806/ej.17.1.200

McKain, M. R., Tang, H., McNeal, J. R., Ayyampalayam, S., Davis, J. I., dePamphilis, C. W., et al. (2016). A phylogenomic assessment of ancient polyploidy and genome evolution across the Poales. *Genome Biol. Evol.* 8, 1150–1164. doi: 10.1093/gbe/evw060

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., et al. (2020). IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* 37, 1530–1534. doi: 10.1093/molbev/msaa015

Murat, F., Xu, J.-H., Tannier, E., Abrouk, M., Guilhot, N., Pont, C., et al. (2010). Ancestral grass karyotype reconstruction unravels new mechanisms of genome shuffling as a source of plant evolution. *Genome Res.* 20, 1545–1557. doi: 10.1101/gr.109744.110

Newell, T. K. (1969). A study of the genus *Joinvillea* (Flagellariaceae). *J. Arnold Arbor.* 50, 527–555.

Ou, S., Chen, J., and Jiang, N. (2018). Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Res.* 46:e126. doi: 10.1093/nar/gky730

Ou, S., and Jiang, N. (2018). LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol.* 176, 1410–1422. doi: 10.1104/pp.17.01310

Ou, S., Su, W., Liao, Y., Chougule, K., Agda, J. R. A., Hellinga, A. J., et al. (2019). Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* 20:275. doi: 10.1186/s13059-019-1905-y

Ouyang, S., Zhu, W., Hamilton, J., Lin, H., Campbell, M., Childs, K., et al. (2007). The TIGR Rice Genome Annotation Resource: improvements and new features. *Nucleic Acids Res.* 35, D883–D887. doi: 10.1093/nar/gkl976

Patel, P., Mathioni, S. M., Hammond, R., Harkess, A. E., Kakrana, A., Arikit, S., et al. (2021). Reproductive phasiRNA loci and *DICER-LIKE5*, but not microRNA loci, diversified in monocotyledonous plants. *Plant Physiol.* 185, 1764–1782. doi: 10.1093/plphys/kiab001

Paterson, A. H., Bowers, J. E., and Chapman, B. A. (2004). Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proc. Natl. Acad. Sci. U.S.A.* 101, 9903–9908. doi: 10.1073/pnas.0307901101

Pohl, R. W., and Davidse, G. (1971). Chromosome numbers of Costa Rican grasses. *Brittonia* 23:293. doi: 10.2307/2805632

Potter, S. C., Luciani, A., Eddy, S. R., Park, Y., Lopez, R., and Finn, R. D. (2018). HMMER web server: 2018 update. *Nucleic Acids Res.* 46, W200–W204. doi: 10.1093/nar/gky448

Preston, J. C., Christensen, A., Malcomber, S. T., and Kellogg, E. A. (2009). MADS-box gene expression and implications for developmental origins of the grass spikelet. *Am. J. Bot.* 96, 1419–1429. doi: 10.3732/ajb.0900062

Preston, J. C., and Kellogg, E. A. (2006). Reconstructing the evolutionary history of paralogous *APETALA1/FRUITFULL-like* genes in grasses (Poaceae). *Genetics* 174, 421–437. doi: 10.1534/genetics.106.057125

Pryszcz, L. P., and Gabaldón, T. (2016). Redundans: an assembly pipeline for highly heterozygous genomes. *Nucleic Acids Res.* 44:e113. doi: 10.1093/nar/gkw294

Rhie, A., Walenz, B. P., Koren, S., and Phillippy, A. M. (2020). Merqury: reference-free quality completeness, and phasing assessment for genome assemblies. *Genome Biol.* 21:245. doi: 10.1186/s13059-020-02134-9

Rodriguez Milla, M. A., Maurer, A., Rodriguez Huete, A., and Gustafson, J. P. (2003). Glutathione peroxidase genes in Arabidopsis are ubiquitous and regulated by abiotic stresses through diverse signaling pathways. *Plant J.* 36, 602–615. doi: 10.1046/j.1365-313X.2003.01901.x

Rothfels, C. J. (2021). Polyploid phylogenetics. *New Phytol.* 230, 66–72. doi: 10.1111/nph.17105

Saarela, J. M., Burke, S. V., Wysocki, W. P., Barrett, M. D., Clark, L. G., Craine, J. M., et al. (2018). A 250 plastome phylogeny of the grass family (Poaceae): topological support under different data partitions. *PeerJ* 6:e4299. doi: 10.7717/peerj.4299

Sajo, M. D. G., Furness, C. A., and Rudall, P. J. (2009). Microsporogenesis is simultaneous in the early-divergent grass *Streptochaeta*, but successive in the closest grass relative, *Ecdeiocolea*. *Grana* 48, 27–37. doi: 10.1080/00173130902746466

Sajo, M. G., Longhi-Wagner, H., and Rudall, P. J. (2007). Floral development and embryology in the early-divergent grass *Pharus*. *Int. J. Plant Sci.* 168, 181–191. doi: 10.1086/509790

Sajo, M. G., Longhi-Wagner, H. M., and Rudall, P. J. (2008). Reproductive morphology of the early-divergent grass *Streptochaeta* and its bearing on the homologies of the grass spikelet. *Plant Syst. Evol.* 275:245. doi: 10.1007/s00606-008-0080-5

Sajo, M. G., Pabón-Mora, N., Jardim, J., Stevenson, D. W., and Rudall, P. J. (2012). Homologies of the flower and inflorescence in the early-divergent grass *Anomochloa* (Poaceae). *Am. J. Bot.* 99, 614–628. doi: 10.3732/ajb.1100290

Salse, J., Bolot, S., Throude, M., Jouffe, V., Piegu, B., Quraishi, U. M., et al. (2008). Identification and characterization of shared duplications between rice and wheat provide new insight into grass genome evolution. *Plant Cell* 20, 11–24. doi: 10.1105/tpc.107.056309

Sarwar, H. (2013). The importance of cereals (Poaceae: Gramineae) nutrition in human health: a review. *J. Cereals Oilseeds* 4, 32–35. doi: 10.5897/JCO12.023

Scanlon, M. J., Schneeberger, R. G., and Freeling, M. (1996). The maize mutant narrow sheath fails to establish leaf margin identity in a meristematic domain. *Development* 122, 1683–1691. doi: 10.1242/dev.122.6.1683

Schmidt, R., Schippers, J. H. M., Mieulet, D., Obata, T., Fernie, A. R., Guiderdoni, E., et al. (2013). MULTIPASS, a rice R2R3-type MYB transcription factor, regulates adaptive growth by integrating multiple hormonal pathways. *Plant J.* 76, 258–273. doi: 10.1111/tpj.12286

Seetharam, A., Singh, U., Li, J., Bhandary, P., Arendsee, Z., and Wurtele, E. S. (2019). Maximizing prediction of orphan genes in assembled genomes. *biorxiv [preprint]* doi: 10.1101/2019.12.17.880294

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31, 3210–3212. doi: 10.1093/bioinformatics/btv351

Soderstrom, T. R. (1981). Some evolutionary trends in the Bambusoideae (Poaceae). *Ann. Mo. Bot. Gard.* 68, 15–47.

Soderstrom, T. R., and Ellis, R. P. (1987). "The position of bamboo genera and allies in a system of grass classification," in *Grass Systematics and Evolution*, eds T. R. Soderstrom, K. W. Hilu, C. S. Campbell, and M. E. Barkworth (Washington, DC: Smithsonian Institution Press), 225–238.

Soreng, R. J., Peterson, P. M., Romaschenko, K., Davidse, G., Teisher, J. K., Clark, L. G., et al. (2017). A worldwide phylogenetic classification of the Poaceae (Gramineae) II: An update and a comparison of two 2015 classifications: phylogenetic classification of the grasses II. *J. Syst. Evol.* 55, 259–290. doi: 10.1111/jse.12262

Spriggs, E. L., Christin, P.-A., and Edwards, E. J. (2014). C4 photosynthesis promoted species diversification during the Miocene grassland expansion. *PLoS One* 9:e97722. doi: 10.1371/journal.pone.0097722

Suyama, M., Torrents, D., and Bork, P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* 34, W609–W612. doi: 10.1093/nar/gkl315

Teng, C., Zhang, H., Hammond, R., Huang, K., Meyers, B. C., and Walbot, V. (2020). Dicer-like 5 deficiency confers temperature-sensitive male sterility in maize. *Nat. Commun.* 11:2912. doi: 10.1038/s41467-020-16634-6

Thody, J., Folkes, L., Medina-Calzada, Z., Xu, P., Dalmay, T., and Moulton, V. (2018). PAREsnip2: a tool for high-throughput prediction of small RNA targets from degradome sequencing data using configurable targeting rules. *Nucleic Acids Res.* 46, 8730–8739. doi: 10.1093/nar/gky609

Tsuji, H., Aya, K., Ueguchi-Tanaka, M., Shimada, Y., Nakazono, M., Watanabe, R., et al. (2006). GAMYB controls different sets of genes and is differentially regulated by microRNA in aleurone cells and anthers. *Plant J.* 47, 427–444. doi: 10.1111/j.1365-313X.2006.02795.x

UniProt Consortium (2021). UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* 49, D480–D489.

UpSetR (2021). *UpSetR Github*. Available online at: https://github.com/hms-dbmi/UpSetR (accessed February 21, 2021).

Valencia, J. I. (1962). Los cromosomas de *Streptochaeta spicata* Schrad. (Gramineae). *Darwiniana* 12, 379–383.

VanBuren, R., Man Wai, C., Wang, X., Pardo, J., Yocca, A. E., Wang, H., et al. (2020). Exceptional subgenome stability and functional divergence in the allotetraploid Ethiopian cereal teff. *Nat. Commun.* 11:884. doi: 10.1038/s41467-020-14724-z

Velasco, R., Zharkikh, A., Affourtit, J., Dhingra, A., Cestaro, A., Kalyanaraman, A., et al. (2010). The genome of the domesticated apple (*Malus* × *domestica* Borkh.). *Nat. Genet.* 42, 833–839. doi: 10.1038/ng.654

Venturini, L., Caim, S., Kaithakottil, G. G., Mapleson, D. L., and Swarbreck, D. (2018). Leveraging multiple transcriptome assembly methods for improved gene structure annotation. *Gigascience* 7:giy093. doi: 10.1093/gigascience/giy093

Walbot, V., and Egger, R. L. (2016). Pre-Meiotic anther development: cell fate specification and differentiation. *Annu. Rev. Plant Biol.* 67, 365–395. doi: 10.1146/annurev-arplant-043015-111804

Wang, J., Yu, J., Sun, P., Li, Y., Xia, R., Liu, Y., et al. (2016). Comparative genomics analysis of rice and pineapple contributes to understand the chromosome number reduction and genomic changes in grasses. *Front. Genet.* 7:174. doi: 10.3389/fgene.2016.00174

Wang, X., Shi, X., Hao, B., Ge, S., and Luo, J. (2005). Duplication and DNA segmental loss in the rice genome: implications for diploidization. *New Phytol.* 165, 937–946. doi: 10.1111/j.1469-8137.2004.01293.x

Whipple, C. J. (2017). Grass inflorescence architecture and evolution: the origin of novel signaling centers. *New Phytol.* 216, 367–372. doi: 10.1111/nph.14538

Whipple, C. J., Zanis, M. J., Kellogg, E. A., and Schmidt, R. J. (2007). Conservation of B class gene expression in the second whorl of a basal grass and outgroups links the origin of lodicules and petals. *Proc. Natl. Acad. Sci. U.S.A.* 104, 1081–1086. doi: 10.1073/pnas.0606434104

White, R. P., Murray, S., Rohweder, M., Prince, S. D., and Thompson, K. M. (2000). *Grassland Ecosystems*. Washington, DC: World Resources Institute Washington, DC, USA.

Xia, R., Xu, J., and Meyers, B. C. (2017). The emergence, evolution, and diversification of the miR390-TAS3-ARF pathway in land plants. *Plant Cell* 29, 1232–1247. doi: 10.1105/tpc.17.00185

Yu, Y., Woo, M.-O., Rihua, P., and Koh, H.-J. (2020). The *DROOPING LEAF (DR)* gene encoding GDSL esterase is involved in silica deposition in rice (*Oryza sativa* L.). *PLoS One* 15:e0238887. doi: 10.1371/journal.pone.0238887

Zhai, J., Arikit, S., Simon, S. A., Kingham, B. F., and Meyers, B. C. (2014). Rapid construction of parallel analysis of RNA end (PARE) libraries for Illumina sequencing. *Methods* 67, 84–90. doi: 10.1016/j.ymeth.2013.06.025

Zhai, J., Zhang, H., Arikit, S., Huang, K., Nan, G.-L., Walbot, V., et al. (2015). Spatiotemporally dynamic, cell-type-dependent premeiotic and meiotic phasiRNAs in maize anthers. *Proc. Natl. Acad. Sci. U.S.A.* 112, 3146–3151. doi: 10.1073/pnas.1418918112

Zhang, H., Liang, W., Yang, X., Luo, X., Jiang, N., Ma, H., et al. (2010). Carbon starved anther encodes a MYB domain protein that regulates sugar partitioning required for rice pollen development. *Plant Cell* 22, 672–689. doi: 10.1105/tpc.109.073668

Zhang, R., Huang, S., Li, S., Song, G., Li, Y., Li, W., et al. (2020). Evolution of PHAS loci in the young spike of allohexaploid wheat. *BMC Genomics* 21:200. doi: 10.1186/s12864-020-6582-4

Zhang, R.-G., Wang, Z.-X., Ou, S., and Li, G.-Y. (2019). TEsorter: lineage-level classification of transposable elements using conserved protein domains. *Biorxiv [preprint]* doi: 10.1101/800177

Zhou, Y., Lu, D., Li, C., Luo, J., Zhu, B.-F., Zhu, J., et al. (2012). Genetic control of seed shattering in rice by the APETALA2 transcription factor SHATTERING ABORTION1. *Plant Cell* 24, 1034–1048. doi: 10.1105/tpc.111.094383

Zhu, J., Chen, H., Li, H., Gao, J.-F., Jiang, H., Wang, C., et al. (2008). *Defective in Tapetal Development and Function 1* is essential for anther development and tapetal function for microspore maturation in Arabidopsis. *Plant J.* 55, 266–277. doi: 10.1111/j.1365-313X.2008.03500.x

Zimin, A. V., Marçais, G., Puiu, D., Roberts, M., Salzberg, S. L., and Yorke, J. A. (2013). The MaSuRCA genome assembler. *Bioinformatics* 29, 2669–2677. doi: 10.1093/bioinformatics/btt476

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.