IsoSeq transcriptome assembly of C_3 panicoid grasses provides tools to study evolutionary change in the Panicoideae

Daniel S. Carvalho[1], Aime V. Nishimwe[1] & James C. Schnable[1]

1. Center for Plant Science Innovation, Department of Agronomy and Horticulture, University of Nebraska-Lincoln, Lincoln, NE, USA.

Abstract

The number of plant species with genomic and transcriptomic data has been increasing rapidly. The grasses – Poaceae – have been well represented among species with published reference genomes. However, as a result the genomes of wild grasses are less frequently targeted by sequencing efforts. Sequence data from wild relatives of crop species in the grasses can aid the study of domestication, gene discovery for breeding and crop improvement, and improve our understanding of the evolution of C_4 photosynthesis. Here we used long read sequencing technology to characterize the transcriptomes of three C_3 panicoid grass species: *Dichanthelium oligosanthes*, *Chasmanthium laxum*, and *Hymenachne amplexicaulis*. Based on alignments to the sorghum genome we estimate that assembled consensus transcripts from each species capture between 54.2 and 65.7% of the conserved syntenic gene space in grasses. Genes co-opted into C_4 were also well represented in this dataset, despite concerns that, because these genes might play roles unrelated to photosynthesis in the target species, they would be expressed at low levels and missed by transcript-based sequencing. A combined analysis using syntenic orthologous genes from grasses with published reference genomes and consensus long read sequences from these wild species was consistent with previously published phylogenies. It is hoped that this data, targeting under represented classes of species within the PACMAD grasses – wild species and species utilizing C_3 photosynthesis – will aid in future studies of domestication and C_4 evolution by decreasing the evolutionary distance between C_4 and C_3 species within this clade, enabling more accurate comparisons associated with evolution of the C_4 pathway.

Introduction

The pace of plant genome sequencing has accelerated in recent years. However, despite decreases in sequencing costs and improvements in genome assembly quality, species selected for whole genome sequencing often meet one or more of the following criteria: A) agricultural importance, B) status as a genetic model system or C) ecological importance. Sequence data from species which lack direct economic, ecological, or genetic model importance can enable comparative analyses to address biological questions in crops and model species (Michael and Jackson, 2013; Ellegren, 2014). C_4 photosynthesis has evolved multiple times in the grasses (GPWG II, 2012), making it particularly amenable to study through comparative genetic approaches (Wang et al., 2009; Huang et al., 2016). C_4 photosynthesis requires both substantial biochemical and anatomical changes (Kellogg, 2013). All grasses which utilize the C_4 pathway belong to the PACMAD clade, a group of grass subfamilies and tribes which includes substantial numbers of both C_3 and C_4 species (GPWG II, 2012). Substantial new insights into both the genes involved in producing the biochemical and anatomical changes required for C_4 photosynthesis, as well as the potential function of individual amino acid residues can be obtained from comparative analysis of individual gene families across species utilizing either C_3 or C_4 photosynthesis within the PACMAD clade (Christin et al., 2007, 2015; Moreno-Villena et al., 2017). However, assembling sequence data for a single gene family from a large enough set of species through PCR amplification and individual Sanger sequencing remains a time and labor intensive process.

Many domesticated grasses belong to the PACMAD clade, including such as maize (*Zea mays*), sugar cane (*Saccharum spp.*), sorghum (*Sorghum bicolor*), and foxtail millet (*Setaria italica*). However every domesticated grass in the PACMAD clade with a sequenced genome utilizes one or more variants of the C_4 photosynthetic pathway (Schnable et al., 2009; Garsmeur et al., 2018; Paterson et al., 2009; Bennetzen et al., 2012). As a result, while published whole genome sequence assemblies exist for at least 14 grasses within the PACMAD clade (Table 1), only one of these (*Dichanthelium oligosanthes*, a wild species) (Studer et al., 2016) utilizes C_3 photosynthesis. Long-read sequencing can effectively generate sequence for large numbers of full length cDNAs even in species lacking reference genome assemblies (An et al., 2018; Zhang et al., 2019). One concern with utilizing this technology for

comparative genetic studies is that the higher error rate, particularly the frequencies of insertion and deletion errors, make data from long read based sequencing of non-model species unsuitable for use in comparative evolutionary analyses (Gonzalez-Garay, 2016). However, we previously found that observed synonymous substitution rates calculated from consensus sequences constructed using PacBio IsoSeq pipeline were not elevated relative to a sister lineage where gene sequences were taken from a sanger-based whole genome assembly, indicating sequence data obtained in this manner may indeed be suitable for comparative evolutionary analyses (Yan et al., 2019).

Species	Relevance	C_3/C_4	Genome Publication
Dichanthelium oligosanthes	Wild Species	C_3	Studer et al. (2016)
Eleusine coracana[a]	Grain Crop	C_4	Hittalmani et al. (2017)
Eragrostis tef [a]	Grain Crop	C_4	Cannarozzi et al. (2014); VanBuren et al. (2019)
Miscanthus x giganteus[b]	Biomass Crop	C_4	Swaminathan et al. (2010)
Oropetium thomaeum[a]	Wild Species	C_4	VanBuren et al. (2015, 2018)
Panicum hallii[c]	Wild Species	C_4	Lovell et al. (2018)
Panicum miliaceum[c]	Grain Crop	C_4	Zou et al. (2019)
Panicum virgatum[c]	Biomass Crop	C_4	Casler et al. (2011)
Pennisetum glaucum[c]	Grain Crop	C_4	Varshney et al. (2017b)
Saccharum spp.[b]	Sugar Crop	C_4	Garsmeur et al. (2018)
Setaria italica[c]	Grain Crop	C_4	Bennetzen et al. (2012)
Setaria viridis[c]	Genetic Model	C_4	Brutnell et al. (2010)
Sorghum bicolor[b]	Grain/Biomass/ Sugar Crop	C_4	Paterson et al. (2009)
Zea mays[b]	Grain Crop & Genetic Model	C_4	Schnable et al. (2009)

Table 1: Published reference genomes for grass species within the PACMAD clade. Species sharing a common inferred evolutionary origin of C_4 photosynthesis as reported in (?) are indicated by superscript letters.

Here we report the sequencing and characterization of IsoSeq based transcriptomes for three additional PACMAD grasses, selecting to enable wider scale studies of protein sequence changes associated with the many parallel origins of C_4 photosynthesis within that clade (Figure 1). These species were specifically selected to augment C_3/C_4 comparisons: *Hymenachne amplexicaulis*, *Chasmanthium laxum*, and *D. oligosanthes*. *H. amplexicaulis* is a member of the grass tribe Paspaleae which contains a mixture of C_3 and C_4 species. The Paspaleae are sister to exclusively C_4 clade. This C_4 clade is variously considered to either consider of two tribes the Andropogoneae and Arundinelleae, or a single expanded Andropogoneae including those species otherwise included in the Arundinelleae. In either nomenclature this clade includes both maize and sorghum, two species with extensive genomic, genetic, and phenotypic resources. *H. amplexicaulis* is found in moist habitats and thrives under flooded conditions (Kibbler and Bahnisch, 1999). *Chasmanthium laxum* belongs to the grass tribe Chasmanthieae (7 species). The Chasmanthieae all appear to utilize C_3 photosynthesis (Kellogg, 2015) and are generally placed as early diverging lineage within the Panicoideae, the grass sub-family containing maize, sorghum, sugar cane, miscanthus, switchgrass, foxtail millet, and proso millet (GPWG II, 2012). *C. laxum* can occur in a variety of environments such as: woods, meadows and swamps (Yates, 1966). The final species targeted for transcriptome sequences was *Dichanthelium oligosanthes*. *D. oligosanthes* is the only PACMAD species exclusively utilizing C_3 photosynthesis with a published genome sequence to date (Studer et al., 2016).

It is a member of the grass tribe Paniceae, a group which also includes foxtail millet, proso millet, and switchgrass, but is an outgroup to the MPC C_4 subclade of exclusively C_4-utilizing species within that tribe (Giussani et al., 2001; GPWG II, 2012; Washburn et al., 2015, 2017). As the published *D. oligosanthes* reference genome was constructed utilizing short read sequencing, the inclusion of *D. oligosanthes* provided an opportunity to improve the proportion of genes with full length sequences from this lineage available for comparative analyses. *D. oligosanthes* is present in small glades on

the edge of woods (A.J. Studer, personal communication, April 08, 2019). The placement of *C. laxum* as an outgroup to other panicoid grasses with sequenced reference genomes and *D. oligosanthes* as sister to other members of the Paniceae with sequences reference genomes were recovered in a preliminary analysis of our long read dataset. Support of the placement of *H. amplexicaulis* as a sister group to Andropogoneae (sorghum and maize) was strong but not unambiguous.

Methods

Plant material, RNA extraction, and sequencing

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

Tissue samples were ground in liquid N_2 and then approximately 200mg of powdered tissue was added to 2 µL of TriPure isolation reagent (Roche Life Science, catalog number #11667157001). The RNA samples mixed with TriPure were then separated using chloroform, precipitated using isopropanol, and RNA pellets were washed using 75% ethanol. The samples were air-dried and diluted in RNAsecure (Ambion). Total RNA concentration was measured using a NanoDrop 1000 spectrophotometer and the integrity was assessed based on electrophoresis on a 1% agarose gel. 10 µL of total RNA for each species was shipped to the Duke Center for Genomic and Computational Biology (GCB), Duke University, USA. Concentrations at the time of shipment ranging from 226.07 to 1,374 ng/µL. OD260/280 ratios for RNA samples were of: 1.93 (*H. amplexicaulis*), 1.92 (*C. Laxum*) and 2.03 (*D. oligosanthes*) within the recommended range for IsoSeq library construction of 1.8-2.2 provided by PacBio. One IsoSeq library was constructed per species and each library was sequenced using a single SMRT cell on a PacBio Sequel.

Consensus reads and transcriptome assembly

Two separate sequence datasets were produced per library: full length (FL) transcripts and non full length (NFL) transcripts. A given transcript was considered FL if the sequence read contained both 5' and 3' adapters as well as poly-A tail and are not redundant to other transcripts. The transcripts lacking the poly-A tail or one of the adapters are instead included in the non-full length dataset. Sequence reads from both files were used to assemble consensus transcriptomes using the software pbtranscript to cluster redundant sequences, part of the SMRT pipe package (version 5.1) with default parameters (https://www.pacb.com/wp-content/uploads/SMRT_Tools_Reference_Guide_v600.pdf). For each final consensus transcript, the single longest ORF present within that transcript was selected as the CDS sequence for downstream analyses. Consensus transcripts were obtained from full-length and non-full length transcripts generated from oligo-dT purified mRNAs. As oligo-dT purification can capture the 3' ends of partially fragmented mRNA molecules it was anticipated that some transcripts may be less than full length and missing start codons. Therefore, ORFs were required to include an in frame stop codon but were not required to include an

in-frame "ATG" which may result in additional non-translated codons being appended to the 5' end of the putative CDS but avoids CDS truncation when the 5' end of the sequence was not recovered.

Sequence data from species with published reference genomes

CDS file containing only one primary transcript per gene downloaded from Phytozome 12 (https://phytozome.jgi.doe.gov/pz/portal.html) was used from *Brachypodium distachyon* (International Brachypodium Initiative et al., 2010), *Oryza sativa* (rice) (Yu et al., 2005; Kawahara et al., 2013), *Sorghum bicolor* (sorghum) (Paterson et al., 2009) and *Setaria italica* (foxtail millet) (Bennetzen et al., 2012). CDS sequences for version 2 of the *Oropetium thomaeum* (oropetium) genome (GenomeID 51527) (VanBuren et al., 2018), and the draft *Eragrostis tef* genome (GenomeID 50954) (VanBuren et al., 2019) were downloaded from CoGe (Lyons and Freeling, 2008). CDS sequences for the initial release of the *Pennisetum glaucum* (pearl millet) genome where downloaded from GigaDB. (Varshney et al., 2017b,a). CDS sequences for B73_RefGenV4 of the *Zea mays* (maize) reference genome was retrieved from Ensembl (Jiao et al., 2017). In cases where only a complete set of CDS sequences was released for a given species, we arbitrarily selected the longest annotated transcript from a given locus to be the single representative transcript for downstream analyses. mRNA sequences from *D. oligosanthes* were obtained CoGe (Genome ID 35847) (Studer et al., 2016) for comparison with IsoSeq *D. oligosanthes* transcripts. PEPC and PPDK gene families from sorghum for further manual curation and phylogeny analysis were obtained from Christin et al. (2007) and Wang et al. (2009), respectively.

Putative orthology assignments

CDS sequences obtained from *H. amplexicaulis*, *C. laxum* and *D. oligosanthes* as described above were compared to the primary CDS sequences of each annotated gene in the sorghum genome using LASTZ version 1.04.00 (Harris, 2007) with the following parameters: –identity=70 –coverage=50 –ambiguous=iupac, –notransition, and –seed=match12. CDS sequences from the three target species were presumed to belong to an orthologous group as a given sorghum gene if the sorghum CDS sequence and target species CDS sequence were reciprocally identified as each others high scoring hit in the LASTZ analysis. The comparison of D. oligosanthes consensus transcripts and annotated mRNA sequences from the published reference genome for D. oligosanthes defined equivalence as sequences which were reciprocally identified as highest scoring LASTZ matches.

Orthologous relationships between sorghum genes and genes in other species with sequenced reference genomes were inferred based on syntenic orthology. For each combination of sorghum and rice, brachypodium, oropetium, teff, foxtail millet, pearl millet, sorghum, and maize all by all LASTZ comparisons were performed using the same parameters described above. The resulting LASTZ output was employed to identify initial syntenic genomic blocks using QuotaAlign with the parameters –tandemNmax=10, cscore=0.5, –merge and –Dm=20 (Tang et al., 2011). The quota was set to –quota=1:2 for maize and teff, and –quota=1:1 for all other species. Pairwise syntenic block data was merged and polished using the methodology previously described in (Zhang et al., 2017) to obtain the final set of high confidence syntenic ortholog groups employed for all downstream analysis.

Orthology was treated as a transitive property, thus each *H. amplexicaulis*, *C. laxum* or *D. oligosanthes* gene identified as putatively orthologous to a given sorghum gene based on reciprocal best LASTZ hit analysis, was also considered to be putatively orthologous to syntenic orthologs of that sorghum gene identified in each of the other species described above. The final sets of putatively orthologous gene groups including both sequences from published reference genomes and the long read sequencing described here is provided in Supplemental Material 1.

Sequence alignment, QC, and phylogenetic analysis

Kalign (v2.04) was used create a multiple sequence alignment from protein sequences obtained by translating CDS sequences from all genes in a give putatively orthologous gene group. This gapped protein alignment was in turn employed to create a codon-level DNA alignment of the original CDS sequences. GBlocks version 0.91 was run with default parameters to identify high quality portions of the sequence alignment and remove those portions of the alignment not meeting specified quality thresholds (Talavera

and Castresana, 2007). Alignments including only those portions passing GBlocks filering were then used as input for RAxML version 8, using the GTRGAMMA model and with a clade of rice and brachypodium specified as an outgroup, to obtain a phylogenetic tree for each group of putatively orthologous genes (Stamatakis, 2014). When RAxML was unable to construct a phylogeny in which rice and brachypodium formed monophyletic clade sister to other other taxa the trees were omitted from downstream visualization. To plot all phylogenies, we used Densitree, part of the BEAST2 package, was used to create combined blots of large numbers of trees (Bouckaert et al., 2014). We performed bootstrap analyses of 100 randomly chosen trees (from all trees generated without any filtering step) with 100 replicates to obtain the branch support values of the most common tree topologies of the Densitree plot. For the trees built for PPDK and PEPC, RAxML bootstrap analyses were performed with 1000 replicates. For visualization purposes only, all branches were treated as having equal length in order to improve the ease of visually comparing differences in topology. Data on the consistency or inconsistency of individual portions of the phylogeny was judged from comparisons of the single best trees generated using the sequence of each separate gene. However, it should be noted that bootstrapping was not performed for the individual best trees for each gene, hence some disagreements in topology may simply reflect poorly resolved nodes with limited support.

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

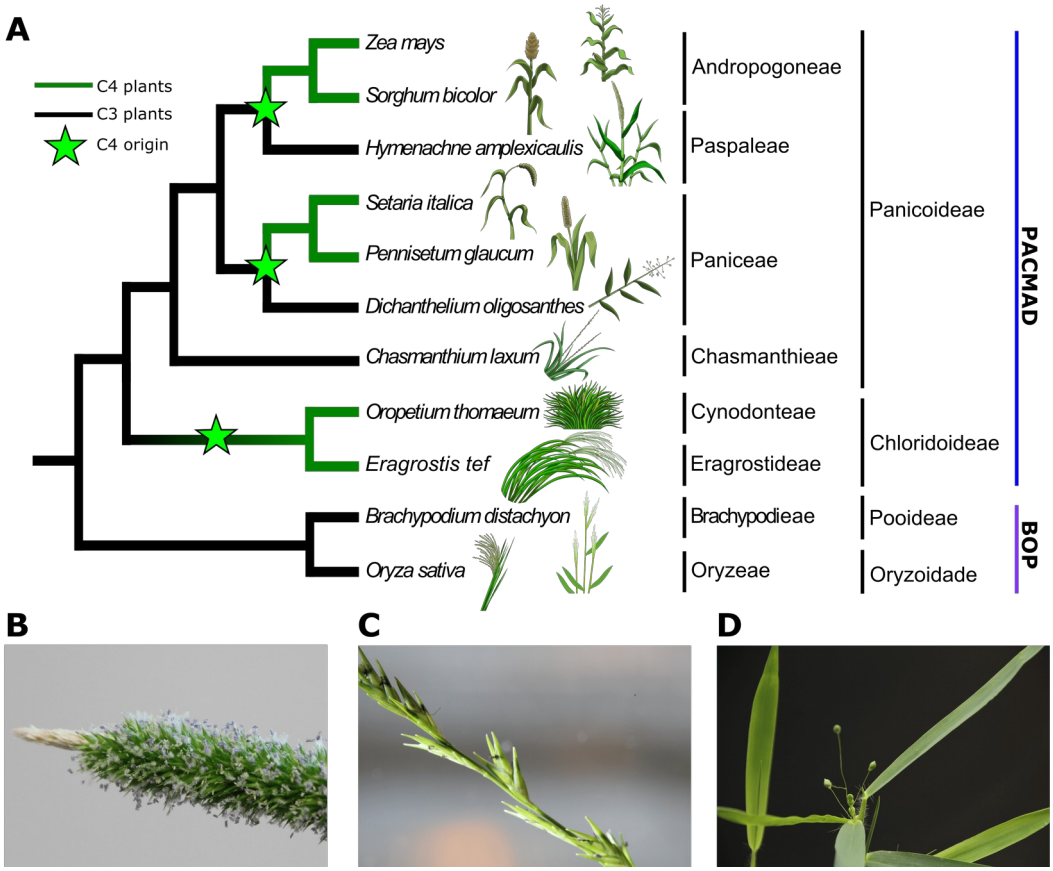

Figure 1: A) Current literature consensus phylogeny of the relationships between the grass species studied here. Lineages in green utilize C_4 photosynthesis, while lineages in black utilize C_3 photosynthesis. The green stars indicate apparent independent origins of C_4 photosynthesis. B) Inflorescence of *H. apmlexicaulis*. C) Inflorescence of *C. laxum*. D) Inflorescence of *D. oligosanthes*.

Results and Discussion

The number of raw reads generated per species was largely consistent and ranged from 708,681 to 734,932 (Table 2). After clustering both full length and non-full length transcripts to obtain a set of polished consensus transcripts, the number of sequences per species dropped to 164,640 to 193,422 (Table 2). The average length of consensus sequences ranged from 925 bp to 1,438 kb (Figure S1). A comparison between the mRNA sequences and IsoSeq transcripts from *D. oligosanthes* was performed to assess the improvement of sequence coverage. Out of the 13,847 reciprocal best LASTZ hits between the mRNA and IsoSeq data, 12,347 transcripts were longer than the mRNA sequences, while the remaining 1,500 sequences were either shorter or the same length in both datasets (Figure S2). The longer sequences recovered in this experiment may be the result of a combination of fragmentation within the *D. oligosanthes* reference genome and improved capture of 5' and 3' untranslated regions, which are often missed by homology-based annotation of genomic sequence. While *H. amplexicaulis* exhibited the shortest consensus transcript length, this was not reflected in a reduced number of complete ORFs – those containing both an in frame ATG and stop codon and occupying at least 60% of the total transcript length. The number of consensus transcripts significantly exceeded the expected number of expressed genes, however, this is consistent with other reference genome-free IsoSeq analyses (Li et al., 2017; Kuang et al., 2019; Yan et al., 2019). Inflated numbers of consensus transcripts can result from sequencing of multiple alternatively spliced isoforms of the same gene, sequencing of incompletely processed mRNA molecules (Martin et al., 2014), high sequence error rates preventing multiple sequences from the same transcript being collapsed into a consensus, divergent haplotypes of the same locus present in our clonally propagated, wild collected, or

partially inbred starting material, or contamination of the original samples with mRNA from non-target organisms.

Species	Total reads	Raw data	CCS reads	FL reads	Average FL length	Consensus transcripts	Average consensus transcript length	Transcripts containing start codon
Hymenachne amplexicaulis	734,932 reads	5.8 GB	732,158 reads	284,027 reads	963 bp	193,422	925 bp	34,016/193,422 (17.5%)
Dichanthelium oligosanthes	708,681 reads	10.1 GB	701,802 reads	380,381 reads	1,460 bp	190,632	1,438 bp	6,055/190,632 (18.9%)
Chasmanthium laxum	729,710 reads	12.5 GB	649,149 reads	306,566 reads	1,294 bp	164,640	1,236 bp	

Table 2: Summary statistics for raw and processed long read sequence data generated from each of the three target species.

Alignment of final consensus reads to the sorghum reference genome was employed to estimate coverage of the shared grass gene space for data collected from each of our target species, as well as to assist in further collapsing multiple redundant sequences originating from alternative splicing, incomplete processing, or divergent haplotypes of transcripts originating from a single genetic locus. In all three cases the majority of consensus transcripts could be aligned to known genes in the sorghum genome, with an average of between 9.3 and 12.1 consensus transcripts aligning to each sorghum gene represented in the transcriptome data (Table 3). Each of these three target species is predicted to be diploid based on either flow cytometry based estimates of genome size and/or imaging of chromosomes, thus a maximum of two transcripts per locus can be explained by divergent haplotypes. The high number of consensus sequences aligned per represented sorghum locus suggests that a large proportion of the overall inflation in consensus transcript number from this dataset may result from alternative splice isoforms or sequencing of incompletely processed mRNA molecules. It should also be noted that this analysis will confound lineage specific gene duplications with divergent haplotypes and splice isoforms, however this bias will be consistent across all three species.

Species	Sorghum gene space coverage	Sorghum syntenic gene space coverage	Transcript alignment rate
H. amplexicaulis	11,485 genes/34,211 genes (33.5%)	6,402 transcripts/11,800 genes (54.2%)	115,361 transcripts/193,422 transcripts (59.6%)
D. oligosanthes	14,159 genes/34,211 genes (41.3%)	7,760 transcripts/11,800 genes (65.7%)	171,465 transcripts/190,632 transcripts (89.9%)
C. laxum	13,446 genes/34,211 genes (39.3%)	7,418 transcripts/11,800 genes (62.8%)	125,357 transcripts/164,640 transcripts (76.1%)

Table 3: Alignment rates of consensus transcripts generated from each of the three target species to the sorghum gene space.

For each sorghum gene which aligned to two or more consensus transcripts from the same target species, a single representative transcript was selected for further downstream analysis (See Methods). Between 11,485 and 14,159 sorghum genes had a corresponding representative transcript in a given target species (Table 3). Here we were: 1) using a single library constructed per species, rather than multiple libraries constructed using different size fractions, 2) using RNA from a single tissue rather than pooled RNA from multiple tissue types, and 3) conducting comparisons between more distantly related species. However, the total proportion of sorghum genes represented in each transcriptome dataset was not substantially lower than the 14,401 *T. dactyloides*-maize gene pairs identified in a previous study which implemented all of these best practices (Yan et al., 2019). This may in part be explained by both sequencing and library preparation improvements between the RSII and Sequel iterations of this sequencing technology.

Manual curation was used to access the coverage and quality of sequences retrieved from these three C_3 photosynthesis-utilizing PACMAD species for five genes known to be involved in C_4 photosynthesis: PPDK, PEPC, NADP-MDH, NAD-ME and DCT2 in C_4 photosynthesis-utilizing PACMAD species. In four cases, the representative transcript identified from each of the three target species spanned every annotated codon in sorghum. The one exception was PPDK where the representative transcript identified for *H. amplexicaulis* lacked the first annotated exon of the annotated gene model in sorghum (Figure 2). Multiple isoforms of the PPDK gene have been described in both maize and sorghum, with the shorter isoform, lacking the same exon absent in *H. amplexicaulis* (Sheen, 1991; Wang et al., 2009). This shorter isoform lacks the chloroplast transit peptide and encodes cytosolic PPDK protein not thought to be associated with C_4 photosynthesis (Glackin and Grula, 1990; Sheen, 1991; Wang et al., 2009). However, these results would also be consistent with sequencing of an incomplete transcript from *H. amplexicaulis* with break point occurring at the same location as the 3' junction of the first exon.

Phylogenetic consistency was assessed using a small subset of genes with high confidence syntenic orthologs identified in species with published reference genomes and representative transcripts identified in each of the three target species. A subset of PACMAD species with sequenced reference genomes were

included in these analyses. Excluded species included those with fragmented genome assemblies at the time of this analyses as well as many, but not all, species with independent whole genome duplications (i.e. *Panicum virgatum*, *Miscanthus x giganteus* and *Eleusine coracana*) as these increased the complexity of downstream analyses. Two well characterized C_4 related proteins PPDK and PEPC were evaluated in detail. PPDK is a member of a small gene family, while genes encoding PEPC proteins are more numerous in many grass species. Copies of each of these two genes in sorghum previously identified as involved in the C_4 cycle were identified from the literature Christin et al. (2007); Wang et al. (2009). Phylogenies constructed using known gene copies from species with published reference genomes and isoseq transcripts of both PPDK and PEPC clustered all C_4 gene copies together (Figure S3). A total of 11,800 genes were identified at syntenic orthologous locations across the genomes of rice, brachypodium, teff, Oropetium, pearl millet, foxtail millet, sorghum, and maize. Of these in 2,774 cases no representative transcripts were retrieved from *C. laxum*, *D. oligosanthes*, or *H. amplexicaulis*. These cases likely represent conserved genes that are not expressed in developing photosynthetic tissue. In 1,611 cases, a representative transcript was identified in only one of the three target species, and in 2,276 cases, representative transcripts were identified in two of the three target species. In the remaining 5,139 cases representative transcripts were retrieved for all three target species. The complete lists of each of these sets of conserved syntenic genes and corresponding transcripts from 0, 1, 2, or 3 of the target species is provided as part of Supplemental Material 1.

One potential concern is using transcriptome data from species utilizing C_3 photosynthesis to provide sequence data for comparative genetic and evolutionary analyses of C_4 is that enzymes involved in the C_4 cycle will likely different functions unrelated to photosynthesis in C_3 plants (Aubry et al., 2011), and therefore may not be expressed in photosynthetic tissue and hence be missing from from datasets derived from sequencing cDNAs. Of 31 core C_4 genes enumerated in (Huang et al., 2016), 20 were part of the set of 11,800 sorghum genes with conserved syntenic orthologs identified in each of the tested grass species with a published reference genome. Hence, these genes are almost certainly present within the genomes of *C. laxum*, *D. oligosanthes*, and *H. amplexicaulis* as well, whether or not they were expressed to sufficient levels to be detected in this analysis. Of these 20 syntenically-conserved C_4 related genes, sequence data was obtained from all three target C_3 utilizing panicoid species in 16 cases. In the remaining four cases – DCT4c, GLR, NADP-ME and SCL – no putatively orthologous transcript was identified in any of the three species. There were no cases where a syntenically conserved gene linked to C_4 photosynthesis was detected in some, but not all, of the three C_3 utilizing species evaluated.

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

The remaining set of 4,908 conserved orthologous gene groups were used to generate protein guided codon multiple sequence alignments (See Methods). A subset of these alignments containing at least 900 nucleotides (300 codons) alignment scored as "high quality" by GBlocks were employed to construct individual gene-level trees (Figure S4). In total 733 trees, representing 267 putatively orthologous gene groups were constructed. Multiple trees resulted from retained duplicate gene pairs resulting from lineage specific whole genome duplications in maize and teff. Each duplication had the potential to create a retained syntenic gene pair which were each co-orthologous to single gene copies in other grass species within the analysis. In order to maintain a consistent number of final nodes, when a retained gene pair was observed in one or both species, multiple sampled trees were generated (see Methods). A modest bias towards over representation of retained – rather than fractionated – genes was observed in the set of genes which were represented in the transcriptome assemblies from all three target species: 37% (1,843/4,908) of maize genes in this set were retained as duplicate pairs vs 30% of all syntenic maize genes, and 100% of teff genes in this set were retained as duplicate pairs vs 91% of all syntenic teff genes. Dating multiple whole genome duplication events

back to the Rho event in grasses is complex (McKain et al., 2016). Therefore, the lack of consideration of these events back to the Rho event could cause a limitation/ambiguity in the gene trees. The rice and brachypodium clade represented a known outgroup as these two species belong to the BOP clade which diverged from the PACMAD clade of grasses early in the evolution of this family (GPWG II, 2012). In 49 cases, RAxML was unable to place the rice brachypodium clade as an outgroup suggested broader issues with orthology assignment, correct ORF identification, or alignment. These trees were not included in downstream analyses.

Among the 684 remaining gene trees, 291 (42.5%) produced a single topology consistent with the prior literature on the relationship of these species (Figure 3). The second and third most common topologies were each represented by less than 7% of all calculated trees, 43 and 28 cases respectively.

The second and third most common topologies differed from prior published phylogenies regarding the placement *H. amplexicaulis*. In the second most common topology *H. amplexicaulis* was placed sister to all other panicoid grass species other than *C. laxum*. In the third most common topology *H. amplexicaulis* was placed sister to the Paniceae. Parallel analysis was conducted using all 4,908 conserved orthologous gene groups, including many cases with substantially shorter regions of high quality multiple sequence alignment. The pattern of trees recovered were largely consistent with those in (Figure 3). In the "all genes" analysis, the same first and second most common topologies were retrieved as in the long alignment only analysis. The third most common in the "all genes" topology places *C. laxum* as sister to the combined Chloridoideae and Panicoideae (Supplemental Figure S5).

A bootstrap analysis was performed and values were retrieved for each of the three most common topologies. In the analysis, 100 randomly trees were chosen to perform the bootstrap analysis (see methods). The most common topology of the 100 trees was the same as the one observed in figure 3, appearing seven times. Both second and third common topologies in figure 3 only appeared once. The bootstrap values observed for the most common topology is the average of all seven bootstrap values of each branch (Figure S6). The bootstrap values of the internal branches in the second and third most common topologies are substantially smaller than the values obtained in the most common topology. Overall, the most common topology exhibits stronger bootstrap support values compared to the other two topologies.

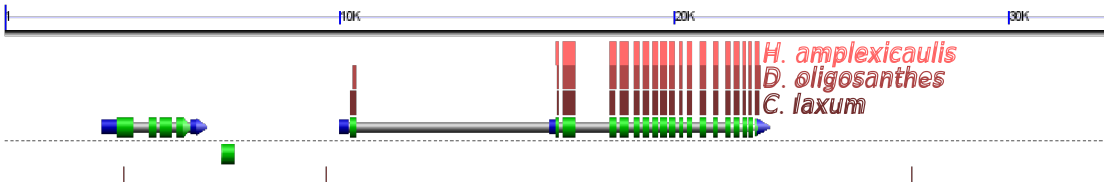

Figure 2: A GEvo panel showing transcript coverage of the C_4 PPDK gene in *S. bicolor* Sobic.009G132900 in each of the three species texted. Red-brown boxes represent regions of similar sequence identified by BLASTN between the sorghum genome and consensus transcript sequences retrieved from *H. amplexicaulis*, *D. oligosanthes*, *C. laxum* (from top most to bottom most). The bottom track indicates the annotated gene structure, with intronic sequence indicated in gray and exonic sequence indicated in either blue (5' or 3' untranslated regions) or green (coding sequence). Top y-axis indicates scale of the displayed genomic region in kilobases. (Lyons and Freeling, 2008).

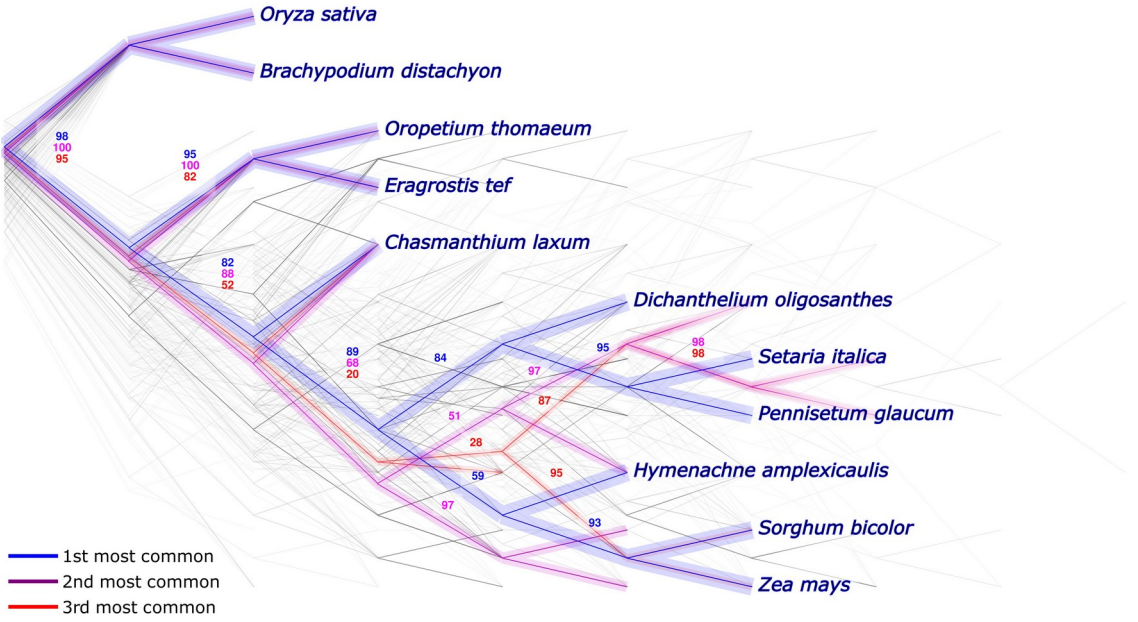

Figure 3: Seven hundred distinct phylogenetic trees calculated from separate multiple sequence alignments of 267 putatively orthologous gene groups with large regions of alignment scored as high quality. Blue indicates the most commonly observed topology (291 trees (42.5% of the total), purple and red indicate the second (43 trees (6.2%) and third most commonly observed topologies (28 trees (4%)), respectively. Numerical labels of branches for each topology indicate average bootstrap support from separately calculated bootstrap trees for 100 randomly selected gene groups, considering data only from those gene trees which supported that particular topology.

Data availability

Raw sequence data for *C. laxum*, *H. amplexicaulis* and *D. oligosanthes* have been deposited in the NCBI SRA under accessions numbers: SRR7632721 (*C. laxum*), SRR7632716 (*H. amplexicaulis*) and SRR9603193 (D. oligosanthes). Processed consensus transcript sequences generated for each of these three species have been deposited in Zenodo DOI 10.5281/zenodo.3253206. Syntenic gene sets and putatively orthologous relationships are provided as Supplemental Material 1.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

DSC and JSC wrote the paper and designed the experiments, DSC generated and analyzed the transcriptome data. All authors have reviewed and approved the manuscript.

Acknowledgements

This work was supported by a CNPq Science Without Borders scholarship (214038/2014-9) to DSC. This material is based upon work supported by the National Science Foundation under Grant No. MCB-183830 to JCS. This project was supported by the Agriculture and Food Research Initiative Grant number 2016-67013-2461 from the USDA National Institute of Food and Agriculture to JCS. In addition, portions of this project were supported by Robert B. Daugherty Water for Food Institute research support award to JCS. The authors would like to thank Lindsay Erndwein who drew the illustrations of various grass species presented in Figure 1A.

Supplemental figures

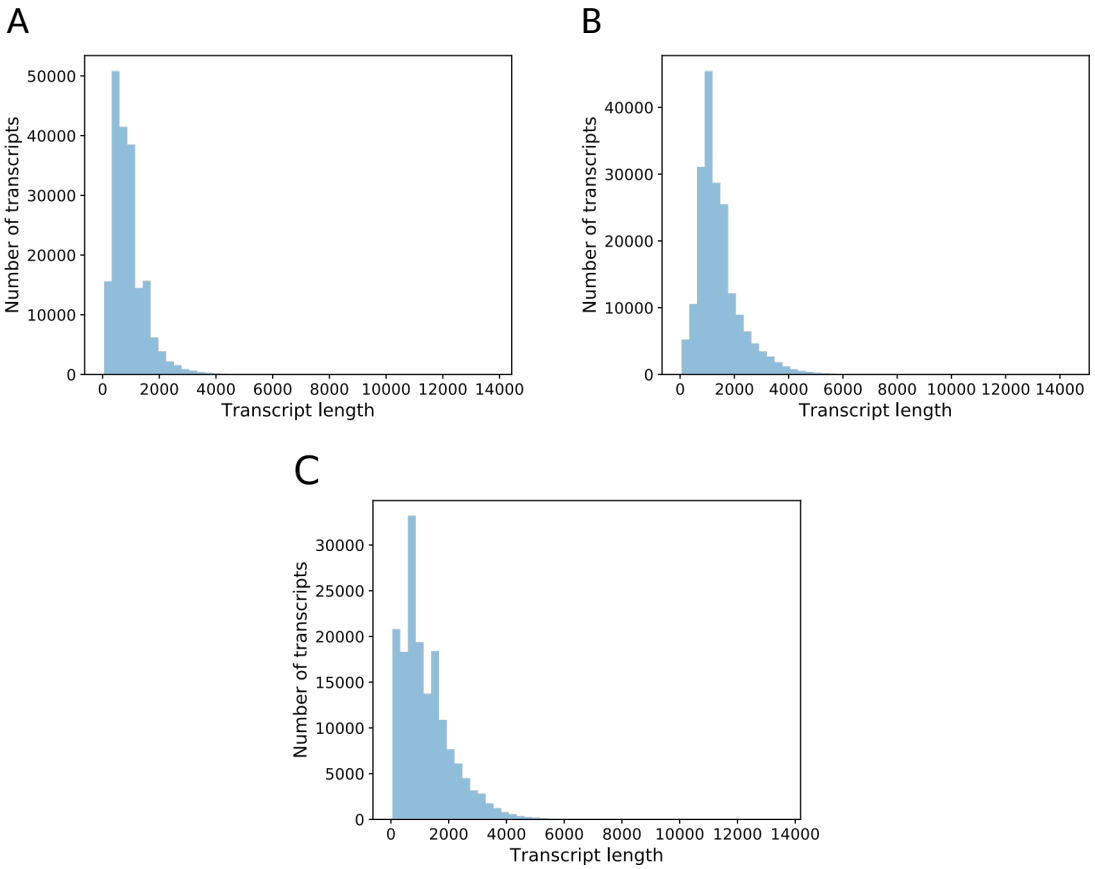

Figure S1: Distribution of lengths for polished transcript sequences: A) *H. amplexicaulis*, B) *D. oligosanthes*, C) *C. laxum*.

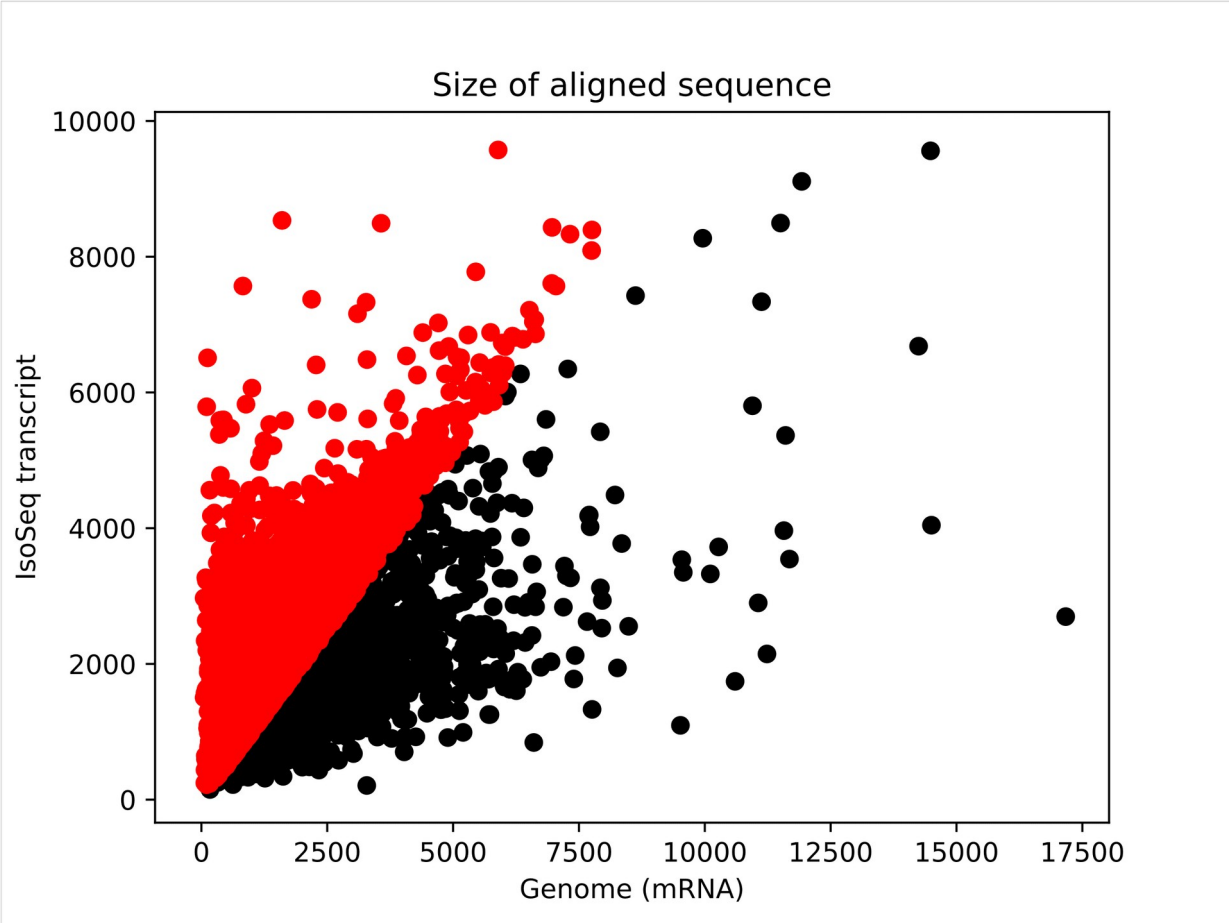

Figure S2: Relationship between aligned mRNA and IsoSeq transcript sizes. Red points indicate sequences where the transcript identified using IsoSeq was longer than the mRNA annotated in the published reference genome for *D. oligosanthes*. Black points indicate sequences where the annotated mRNA sequence was longer than the corresponding sequenced transcript.

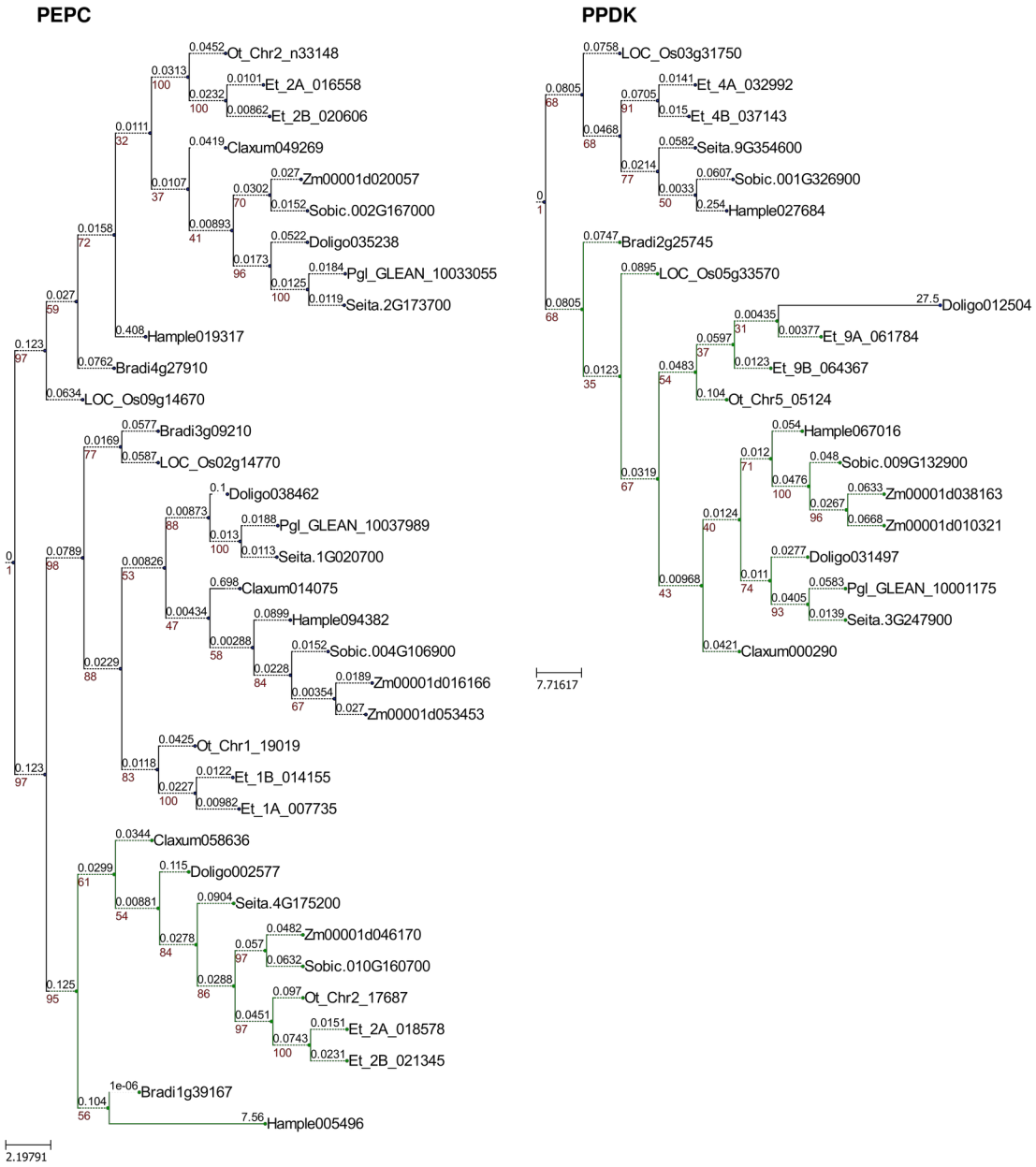

Figure S3: Maximum-likelihood phylogenetic trees of PEPC and PPDK gene families. Green branches represent branches leading to genes from the same orthologous group as the sorghum C_4 gene copy. Branch lengths and bootstrap values are shown in black and red, respectively.

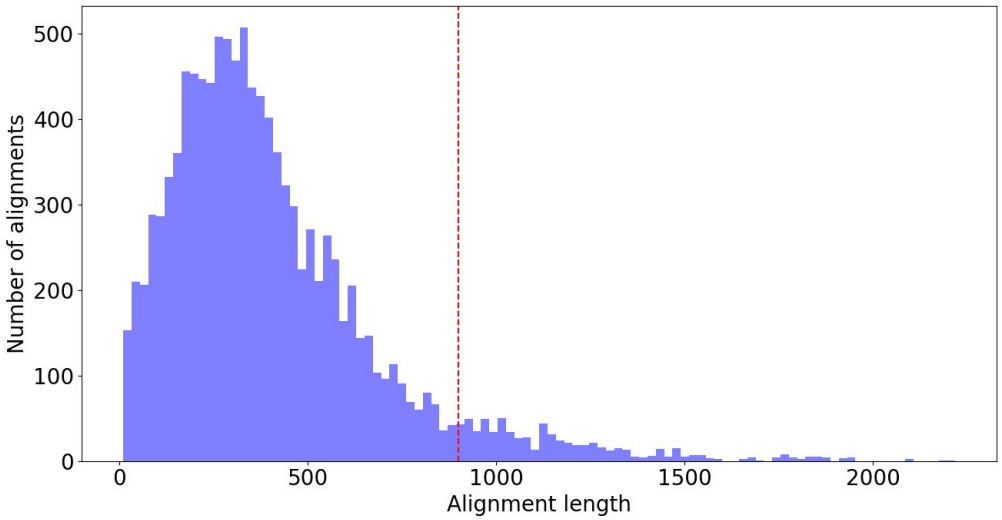

Figure S4: Distribution of alignment lengths after GBlocks cleaning. Dashed line represents the threshold of 900 nucleotides long sequences employed for Figure 3. Sequences represented on the right side of the histogram were analyzed.

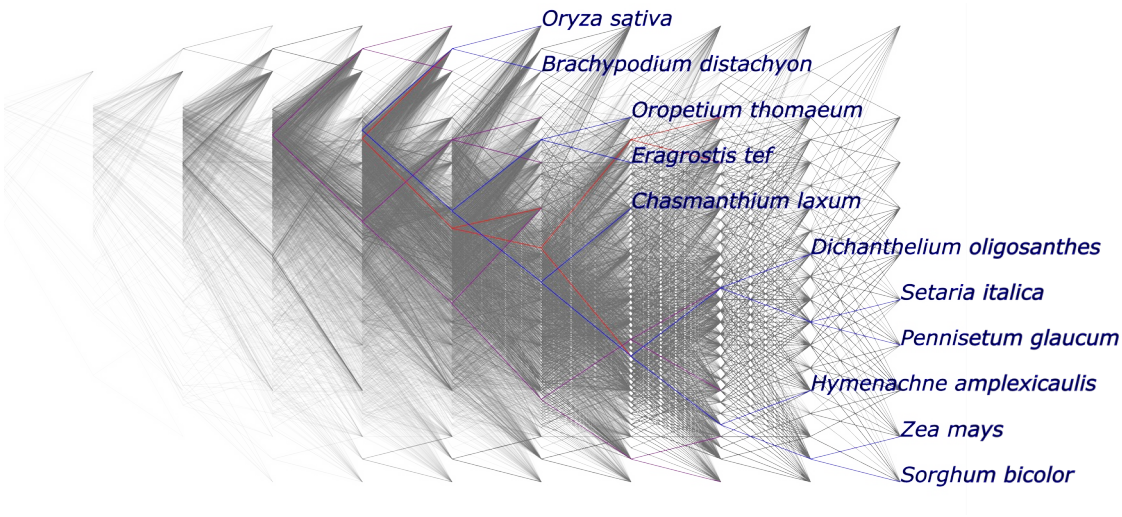

Figure S5: Plot of 11,371 phylogenetic trees. The blue branches represent the most common topology, purple and red branches represent second and third most common topologies, respectively.
Figure generated using Densitree.

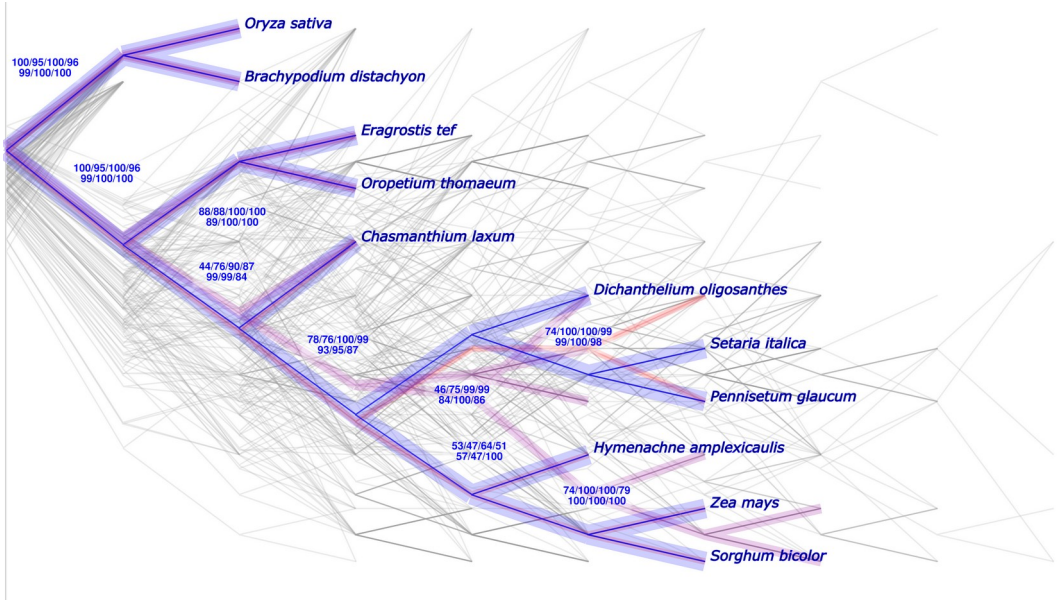

Figure S6: Plot of 100 phylogenetic trees randomly selected. The blue branches represent the most common topology (7 trees), purple and red branches represent the other two most common topologies (2 trees each). Bootstrap values (in blue) obtained from each of the seven trees of the common topology. The bootstrap values in the same position are from the same tree. Figure generated using Densitree.

References

An, D., Cao, H., Li, C., Humbeck, K. and Wang, W. (2018). Isoform sequencing and state-of-art applications for unravelling complexity of plant transcriptomes. *Genes* 9, 43.

Aubry, S., Brown, N. J. and Hibberd, J. M. (2011). The role of proteins in c3 plants prior to their recruitment into the c4 pathway. *Journal of experimental botany* 62, 3049–3059.

Bennetzen, J. L., Schmutz, J., Wang, H., Percifield, R., Hawkins, J., Pontaroli, A. C., Estep, M., Feng, L., Vaughn, J. N., Grimwood, J. et al. (2012). Reference genome sequence of the model plant setaria. *Nature biotechnology* 30, 555.

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., Suchard, M. A., Rambaut, A. and Drummond, A. J. (2014). Beast 2: a software platform for bayesian evolutionary analysis. *PLoS computational biology* 10, e1003537.

Brutnell, T. P., Wang, L., Swartwood, K., Goldschmidt, A., Jackson, D., Zhu, X.-G., Kellogg, E. and Van Eck, J. (2010). Setaria viridis: a model for c4 photosynthesis. *The Plant Cell* 22, 2537–2544.

Cannarozzi, G., Plaza-Wüthrich, S., Esfeld, K., Larti, S., Wilson, Y. S., Girma, D., de Castro, E., Chanyalew, S., Blösch, R., Farinelli, L. et al. (2014). Genome and transcriptome sequencing identifies breeding targets in the orphan crop tef (eragrostis tef). *BMC genomics* 15, 581.

Casler, M. D., Tobias, C. M., Kaeppler, S. M., Buell, C. R., Wang, Z.-Y., Cao, P., Schmutz, J. and Ronald, P. (2011). The switchgrass genome: tools and strategies. *The Plant Genome* 4, 273–282.

Christin, P.-A., Arakaki, M., Osborne, C. P. and Edwards, E. J. (2015). Genetic enablers underlying the clustered evolutionary origins of c4 photosynthesis in angiosperms. *Molecular biology and evolution* 32, 846–858.

Christin, P.-A., Salamin, N., Savolainen, V., Duvall, M. R. and Besnard, G. (2007). C 4 photosynthesis evolved in grasses via parallel adaptive genetic changes. *Current Biology* 17, 1241–1247.

411 Ellegren, H. (2014). Genome sequencing and population genomics in non-model organisms. 412 *Trends in ecology & evolution* 29, 51–63.

413 Garsmeur, O., Droc, G., Antonise, R., Grimwood, J., Potier, B., Aitken, K., Jenkins, 414 J., Martin, G., Charron, C., Hervouet, C. et al. (2018). A mosaic monoploid reference 415 sequence for the highly complex genome of sugarcane. *Nature communications* 9, 2638.

416 Giussani, L. M., Cota-Sánchez, J. H., Zuloaga, F. O. and Kellogg, E. A. (2001). A 417 molecular phylogeny of the grass subfamily panicoideae (poaceae) shows multiple origins of c4 418 photosynthesis. *American Journal of Botany* 88, 1993–2012.

419 Glackin, C. A. and Grula, J. W. (1990). Organ-specific transcripts of different size and 420 abundance derive from the same pyruvate, orthophosphate dikinase gene in maize. *Proceedings* 421 *of the National Academy of Sciences* 87, 3004–3008.

422 Gonzalez-Garay, M. L. (2016). Introduction to isoform sequencing using pacific biosciences 423 technology (iso-seq). In *Transcriptomics and Gene Regulation*, pp. 141–160. Springer. 424 GPWG II (2012). New grass phylogeny resolves deep evolutionary relationships and discovers c4 425 origins. *New Phytologist* 193, 304–312.

426 Harris, R. S. (2007). *Improved pairwise alignment of genomic DNA*. The Pennsylvania State 427 University.

428 Hittalmani, S., Mahesh, H., Shirke, M. D., Biradar, H., Uday, G., Aruna, Y., Lo429 hithaswa, H. and Mohanrao, A. (2017). Genome and transcriptome sequence of finger 430 millet (eleusine coracana (l.) gaertn.) provides insights into drought tolerance and nutraceutical 431 properties. *BMC genomics* 18, 465.

432 Huang, P., Studer, A. J., Schnable, J. C., Kellogg, E. A. and Brutnell, T. P. (2016). 433 Cross species selection scans identify components of c4 photosynthesis in the grasses. *Journal of* 434 *Experimental Botany* 68, 127–135.

435 Initiative, I. B. et al. (2010). Genome sequencing and analysis of the model grass brachypodium 436 distachyon. *Nature* 463, 763.

437 Jiao, Y., Peluso, P., Shi, J., Liang, T., Stitzer, M. C., Wang, B., Campbell, M. S., 438 Stein, J. C., Wei, X., Chin, C.-S. et al. (2017). Improved maize reference genome with 439 single-molecule technologies. *Nature* 546, 524.

440 Kawahara, Y., de la Bastide, M., Hamilton, J. P., Kanamori, H., McCombie, W. R.,

441 Ouyang, S., Schwartz, D. C., Tanaka, T., Wu, J., Zhou, S. et al. (2013). Improvement 442 of the oryza sativa nipponbare reference genome using next generation sequence and optical map 443 data. *Rice* 6, 4.

444 Kellogg, E. A. (2013). C4 photosynthesis. *Current Biology* 23, R594–R599.

445 Kellogg, E. A. (2015). Viii. subfamily panicoideae link (1827). In *Flowering Plants. Monocots*, 446 pp. 271–345. Springer.

447 Kibbler, H. and Bahnisch, L. (1999). Physiological adaptations of hymenachne amplexicaulis 448 to flooding. *Australian journal of experimental agriculture* 39, 429–435.

449 Kuang, X., Sun, S., Wei, J., Li, Y. and Sun, C. (2019). Iso-seq analysis of the taxus 450 cuspidata transcriptome reveals the complexity of taxol biosynthesis. *BMC plant biology* 19, 451 210.

452 Li, J., Harata-Lee, Y., Denton, M. D., Feng, Q., Rathjen, J. R., Qu, Z. and Adelson, 453 D. L. (2017). Long read reference genome-free reconstruction of a full-length transcriptome from 454 astragalus membranaceus reveals transcript variants involved in bioactive compound biosynthe455 sis. *Cell discovery* 3, 17031.

456 Lovell, J. T., Jenkins, J., Lowry, D. B., Mamidi, S., Sreedasyam, A., Weng, X., Barry, 457 K., Bonnette, J., Campitelli, B., Daum, C. et al. (2018). The genomic landscape of 458 molecular responses to natural drought stress in panicum hallii. *Nature communications* 9, 459 5213.

460 Lyons, E. and Freeling, M. (2008). How to usefully compare homologous plant genes and 461 chromosomes as dna sequences. *The Plant Journal* 53, 661–673.

Martin, J. A., Johnson, N. V., Gross, S. M., Schnable, J., Meng, X., Wang, M., Coleman-Derr, D., Lindquist, E., Wei, C.-L., Kaeppler, S. et al. (2014). A near complete snapshot of the zea mays seedling transcriptome revealed from ultra-deep sequencing. *Scientific reports* 4, 4519.

McKain, M. R., Tang, H., McNeal, J. R., Ayyampalayam, S., Davis, J. I., dePamphilis, C. W., Givnish, T. J., Pires, J. C., Stevenson, D. W. and Leebens-Mack, J. H. (2016). A phylogenomic assessment of ancient polyploidy and genome evolution across the poales. *Genome biology and evolution* 8, 1150–1164.

Michael, T. P. and Jackson, S. (2013). The first 50 plant genomes. *The plant genome* 6.

Moreno-Villena, J. J., Dunning, L. T., Osborne, C. P. and Christin, P.-A. (2017). Highly expressed genes are preferentially co-opted for c4 photosynthesis. *Molecular biology and evolution* 35, 94–106.

Paterson, A. H., Bowers, J. E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., Haberer, G., Hellsten, U., Mitros, T., Poliakov, A. et al. (2009). The sorghum bicolor genome and the diversification of grasses. *Nature* 457, 551.

Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T. A. et al. (2009). The b73 maize genome: complexity, diversity, and dynamics. *science* 326, 1112–1115.

Sheen, J. (1991). Molecular mechanisms underlying the differential expression of maize pyruvate, orthophosphate dikinase genes. *The Plant Cell* 3, 225–245.

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

Studer, A. J., Schnable, J. C., Weissmann, S., Kolbe, A. R., McKain, M. R., Shao, Y., Cousins, A. B., Kellogg, E. A. and Brutnell, T. P. (2016). The draft genome of the c 3 panicoid grass species dichanthelium oligosanthes. *Genome biology* 17, 223.

Swaminathan, K., Alabady, M. S., Varala, K., De Paoli, E., Ho, I., Rokhsar, D. S., Arumuganathan, A. K., Ming, R., Green, P. J., Meyers, B. C. et al. (2010). Genomic and small rna sequencing of miscanthus× giganteus shows the utility of sorghum as a reference genome sequence for andropogoneae grasses. *Genome biology* 11, R12.

Talavera, G. and Castresana, J. (2007). Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Systematic biology* 56, 564– 577. Tang, H., Lyons, E., Pedersen, B., Schnable, J. C., Paterson, A. H. and Freeling, M. (2011). Screening synteny blocks in pairwise genome comparisons through integer programming. *BMC bioinformatics* 12, 102.

VanBuren, R., Bryant, D., Edger, P. P., Tang, H., Burgess, D., Challabathula, D., Spittle, K., Hall, R., Gu, J., Lyons, E. et al. (2015). Single-molecule sequencing of the desiccation-tolerant grass oropetium thomaeum. *Nature* 527, 508.

VanBuren, R., Wai, C. M., Keilwagen, J. and Pardo, J. (2018). A chromosome-scale assembly of the model desiccation tolerant grass oropetium thomaeum. *Plant Direct* 2, e00096.

VanBuren, R., Wai, C. M., Pardo, J., Yocca, A. E., Wang, X., Wang, H., Chaluvadi, S. R., Bryant, D., Edger, P. P., Bennetzen, J. L. et al. (2019). Exceptional subgenome stability and functional divergence in allotetraploid teff, the primary cereal crop in ethiopia. *bioRxiv* p. 580720.

Varshney, R., Liu, X., Shi, C., Vigouroux, Y. and Xu, X. (2017a). Genomic data of pearl millet (Pennisetum glaucum). dx.doi.org/10.5524/100192.

Varshney, R. K., Shi, C., Thudi, M., Mariac, C., Wallace, J., Qi, P., Zhang, H., Zhao, Y., Wang, X., Rathore, A. et al. (2017b). Pearl millet genome sequence provides a resource to improve agronomic traits in arid environments. *Nature biotechnology* 35, 969.

Wang, X., Gowik, U., Tang, H., Bowers, J. E., Westhoff, P. and Paterson, A. H. (2009). Comparative genomic analysis of c4 photosynthetic pathway evolution in grasses. *Genome biology* 10, R68.

Washburn, J. D., Schnable, J. C., Conant, G. C., Brutnell, T. P., Shao, Y., Zhang, Y., Ludwig, M., Davidse, G. and Pires, J. C. (2017). Genome-guided phylo-transcriptomic methods and the nuclear phylogenetic tree of the paniceae grasses. *Scientific reports* 7, 13528. Washburn, J. D., Schnable, J. C., Davidse, G. and Pires, J. C. (2015). Phylogeny and photosynthesis of the grass tribe paniceae. *American Journal of Botany* 102, 1493–1505.

Yan, L., Kenchanmane Raju, S. K., Lai, X., Zhang, Y., Dai, X., Rodriguez, O., Mahboub, S., Roston, R. L. and Schnable, J. C. (2019). Parallels between natural selection in the cold-adapted crop-wild relative tripsacum dactyloides and artificial selection in temperate adapted maize. *The Plant Journal* .

Yates, H. O. (1966). Revision of grasses traditionally referred to uniola, ii. chasmanthium. *The Southwestern Naturalist* pp. 415–455.

Yu, J., Wang, J., Lin, W., Li, S., Li, H., Zhou, J., Ni, P., Dong, W., Hu, S., Zeng, C. et al. (2005). The genomes of oryza sativa: a history of duplications. *PLoS biology* 3, e38.

Zhang, G., Sun, M., Wang, J., Lei, M., Li, C., Zhao, D., Huang, J., Li, W., Li, S., Li, J. et al. (2019). Pacbio full-length cdna sequencing integrated with rna-seq reads drastically improves the discovery of splicing transcripts in rice. *The Plant Journal* 97, 296–305.

Zhang, Y., Ngu, D. W., Carvalho, D., Liang, Z., Qiu, Y., Roston, R. L. and Schnable, J. C. (2017). Differentially regulated orthologs in sorghum and the subgenomes of maize. *The Plant Cell* 29, 1938–1951.

Zou, C., Li, L., Miki, D., Li, D., Tang, Q., Xiao, L., Rajput, S., Deng, P., Peng, L., Jia, W. et al. (2019). The genome of broomcorn millet. *Nature communications* 10, 436.