Published September 20, 2018



# *Tripsacum* De novo Transcriptome Assemblies Reveal Parallel Gene Evolution with Maize after Ancient Polyploidy

Christine M. Gault,* Karl A. Kremling, and Edward S. Buckler

C.M. Gault, Institute of Genomic Diversity, Cornell Univ., 526 Campus Road, Ithaca, NY 14853; K.A. Kremling, Plant Breeding and Genetics Section, Cornell Univ., 526 Campus Road, Ithaca, NY 14853; E.S. Buckler, USDA–ARS; 526 Campus Road, Ithaca, NY, 14853.

ABSTRACT Plant genomes reduce in size following a whole-genome duplication event, and one gene in a duplicate gene pair can lose function in absence of selective pressure to maintain duplicate gene copies. Maize (*Zea mays* L.) and its sister genus, *Tripsacum*, share a genome duplication event that occurred 5 to 26 million years ago. Because few genomic resources for *Tripsacum* exist, it is unknown whether *Tripsacum* grasses and maize have maintained a similar set of genes that have resisted decay into pseudogenes. Here we present high-quality de novo transcriptome assemblies for two species: *Tripsacum dactyloides* (L.) L. and *T. floridanum* Porter ex Vasey. Genes with experimental protein evidence in maize were good candidates for genes resistant to pseudogenization in both genera because pseudogenes by definition do not produce protein. We tested whether 15,160 maize genes with protein evidence are resisting gene loss and whether their *Tripsacum* homologs are also resisting gene loss. Protein-encoding maize transcripts and their *Tripsacum* homologs have higher guanine–cytosine (GC) content, higher gene expression levels, and more conserved expression levels than putatively untranslated maize transcripts and their *Tripsacum* homologs. These results suggest that similar genes may be decaying into pseudogenes in both genera after a shared ancient polyploidy event. The *Tripsacum* transcriptome assemblies provide a high-quality genomic resource that can provide insight into the evolution of maize, a highly valuable crop worldwide.

**Abbreviations:** FPKM, fragments per kilobase of transcript per million mapped reads; GC, guanine–cytosine; GO, gene ontology; PASA, Program to Assemble Spliced Alignments.

## CORE IDEAS

- Maize genes with protein evidence have higher expression and GC content
- *Tripsacum* homologs of maize genes exhibit the same trends as in maize
- Maize proteome genes have more highly correlated gene expression with *Tripsacum*
- Expression dominance for homeologs occurs similarly between maize and *Tripsacum* homologs
- A similar set of genes may be decaying into pseudogenes in maize and *Tripsacum*

Developing genomic resources for wild relatives of crops can aid crop breeding because they often possess novel desirable traits such as resistance to biotic and abiotic stresses. Wild relatives have not passed through a domestication bottleneck and possess untapped genetic diversity. The perennial grass genus *Tripsacum* is the closest sister genus to *Zea* (Bomblies and Doebley, 2005; Mathews et al., 2002), the genus that contains maize. As some of the most ecologically dominant grasses in the Americas, *Tripsacum* grasses possess freezing tolerance and perenniality, which are traits that maize lacks. The

Citation: Gault, C.M., K.A. Kremling, and E.S. Buckler. 2018. Tripsacum De novo Transcriptome Assemblies Reveal Parallel Gene Evolution with Maize after Ancient Polyploidy. Plant Genome 11:180012. doi: 10.3835/plantgenome2018.02.0012

Received 21 Feb. 2018. Accepted 2 July 2018.
*Corresponding author (cg449@cornell.edu).

natural habitat of the *Tripsacum* genus is vast, extending from the northern United States at 42°N latitude to Argentina at 24°S latitude and encompassing both tropical regions and temperate regions with harsh winters (deWet et al., 1982). *Tripsacum* grasses overwinter in subzero temperatures in temperate zones. In contrast, freezing temperatures cause maize leaf damage, yield reduction, or death (Li et al., 2016; Carter, 1995; Elmore and Doupnik, 1995). In addition to abiotic stress tolerance, *T. dactyloides*, also known as Eastern Gamagrass, has resistance to three maize pests: the *Striga hermonthica* (Delile) Benth. parasitic weed (Gurney et al., 2003), western corn rootworm (Branson, 1971; Moellenbeck et al., 1995), and *Puccinia sorghi* Schwein. (common rust) (Bergquist, 1981). The genetic architecture of these beneficial traits has not been fully explored because there are few *Tripsacum* genomic resources. Currently available *Tripsacum* resources include whole-genome skim sequencing datasets for diploid and tetraploid *Tripsacum* lines (Chia et al., 2012; Zhu et al., 2016), but an assembled *Tripsacum* genome has not been published. Additionally, 24,616 *T. dactyloides* isoforms have been assembled and analyzed, revealing that phospholipid biosynthesis genes show rapid evolution in *T. dactyloides* and may have aided temperate adaptation (Yan et al., 2018). Maize, the most productive cereal crop in the world (FAOSTAT, http://faostat.fao.org), could be improved by bringing in valuable traits from its wild relative, *Tripsacum*.

The *Tripsacum* gene complement can frame maize gene evolution in a larger context. The *Tripsacum* genus diverged from the *Zea* genus <1.2 million years ago, which was before maize domestication (Ross-Ibarra et al., 2009). *Tripsacum* and *Zea* make up the subtribe Tripsacinae in tribe Andropogoneae, subfamily Panicoideae, family Poaceae (Soreng et al., 2015). The two genera share a genome duplication event that occurred 5 to 26 million years ago, around the time their shared lineage diverged from the sorghum [*Sorghum bicolor* (L.) Moench] lineage (Wang et al., 2015; Swigonová et al., 2004; Whitkus et al., 1992; Berhan et al., 1993). Immediately after the tetraploidization event, the ancestor of *Tripsacum* and maize had 20 chromosomes that eventually became rearranged through chromosomal breakage and fusion (Wei et al., 2007; Murat et al., 2010). The maize genome has been reorganized into 10 chromosomes, whereas *Tripsacum* species have a base chromosome number of 18 and can be diploid ($2n = 36$), triploid ($2n = 54$), or tetraploid ($2n = 72$) (deWet et al., 1982). Despite the difference in chromosome number, there are no evident large-scale chromosomal deletions between the two genera, and they have very similar gene content (Chia et al., 2012).

The maize and *Tripsacum* genomes are still in flux after the whole-genome duplication event. Duplicate genes arising from this ancient tetrapolyploidy event are called homeologous genes and can have many evolutionary fates. Most commonly, one homeolog loses gene function by pseudogenization or deletion in absence of selective pressure to maintain both homeologs (Lynch and Conery, 2000; Maere et al., 2005). In other cases, both homeologs are retained after polyploidization. Genes with dosage balance sensitivity tend to maintain their prepolyploidy stoichiometric ratio and both homeologs are retained (Tasdighian et al., 2017; Conant et al., 2014). Retained homeologs can undergo subfunctionalization or neofunctionalization (Hughes et al., 2014; Pophaly and Tellier, 2015).

Biased gene fractionation (gene loss) occurs when one homeologous region loses more genes than the other homeologous region (Freeling, 2009). In *Arabidopsis*, 85% of homeologous regions showed biased gene fractionation, where one region was 1.6 times more likely to lose genes than its homeolog (Thomas et al., 2006). In maize, 68% of homeologous regions showed biased gene fractionation, where one region was 2.3 times more likely to lose genes than its homeolog (Woodhouse et al., 2010). Genome dominance occurs when the subgenome contributed by one parent retains more genes and has higher gene expression than the subgenome contributed by the other parent (Woodhouse et al., 2014). Maize homeologous regions were partitioned into the dominant subgenome 1 and the recessive subgenome 2 based on these gene retention and expression criteria (Schnable et al., 2011). Of the homeologous gene pairs that were retained in maize inbred line B73, more genes were lost from subgenome 2 than subgenome 1 in multiple maize inbred lines (Schnable et al., 2011, Brohammer et al., 2018).

The factors controlling genome dominance are not fully elucidated. Garsmeur et al. (2014) postulate that genome dominance may be more likely to occur in species that have experienced allopolyploidy rather than autopolyploidy. Zhao et al. (2017) refine this model and propose that genome dominance occurs when the subgenomes are highly diverged but not when the subgenomes are genetically similar. Supporting this proposal, no subgenome dominance was observed between the closely related soybean [*Glycine max* (L.) Merr.] progenitor genomes, but subgenome dominance was observed between the divergent maize progenitor genomes (Zhao et al., 2017). Species arising from the same whole-genome duplication event behave similarly in the rate of duplicate gene loss and biased fractionation (Sankoff et al., 2010), but any parallels in maize and *Tripsacum* evolution have yet to be studied. Maize and *Tripsacum* have been independently reducing their gene content after their shared ancient tetraploidy event. Approximately 60% of maize homeolog pairs have been reduced to singletons in the maize genome so far (Schnable et al., 2011). Given enough time, more than 90% of gene copies are eventually lost in eukaryotic genomes after whole-genome duplication (Sankoff et al., 2010). Gene loss happens slowly; the half-life for duplicated genes in the *Arabidopsis thaliana (L.) Heynh.* genome is 17.3 million years (Lynch and Conery, 2003). Thus, the maize and *Tripsacum* genomes are still shedding duplicated genes. Understanding which genes are resistant to pseudogenization in *Tripsacum* and maize would greatly aid crop improvement efforts; however, the maize and *Tripsacum* transcriptomes have yet to be compared to identify a

common gene set that is resistant to pseudogenization and gene loss since their shared polyploidy event.

Genes that are resistant to pseudogenization in maize and *Tripsacum* are predicted to be located near recombination hotspots. High recombination rates are associated with the removal of deleterious alleles (Rodgers-Melnick et al., 2015; Tiley et al., 2015). Recombination unlinks beneficial variants from deleterious mutations, which may interrupt the coding sequence and create a pseudogene. Recombination also strongly influences nucleotide composition through GC-biased gene conversion. When recombination occurs at heterozygous sites, a heteroduplex forms. The mismatch repair machinery often favors the G/C allele over the A/T allele, leading to GC-biased gene conversion. The GC-biased gene conversion increases the GC content in all three codon positions as well as introns and intergenic recombination hotspots. The GC-biased gene conversion plays a major role in shaping GC content in maize, rice (*Oryza sativa* L.), and other plant genomes (Rodgers-Melnick et al., 2015; Muyle et al., 2011; Serres-Giardi et al., 2012). Although base composition is shaped by other evolutionary forces such as mutational bias and selection on codon usage, GC-biased gene conversion has emerged as the strongest force affecting base composition (Clément et al., 2017). Thus, genes resistant to gene loss in maize and *Tripsacum* are predicted to have higher recombination rates and higher GC content than other genes in the genome, although this has yet to be tested.

Here, we present high-quality de novo transcriptome assemblies for two *Tripsacum* species: *T. dactyloides* and *T. floridanum*. We hypothesize that a subset of genes shared by *Tripsacum* and maize are resistant to gene loss in both clades. Genes supported by experimental protein evidence in maize are good candidates for genes that are resistant to pseudogenization and gene loss in maize and *Tripsacum* because functional genes produce protein, while pseudogenes cannot produce protein because of disabling mutations affecting the coding sequence (Xiao et al., 2016). A recent study by Walley et al. (2016) found that less than half of maize RefGen_v4 genes (15,160 out of 39,324) are detected in the proteome, which was constructed using electrospray ionization tandem mass spectrometry (Walley et al., 2016). Genes with detected protein tend to be syntenically conserved with sorghum, indicating that they are under high selective pressure (Walley et al., 2016). We tested whether maize genes with protein evidence and their *Tripsacum* homologs have higher GC content, higher expression, and more tightly conserved expression levels than putatively untranslated maize genes and their *Tripsacum* homologs. We also tested whether expression dominance was conserved in homeologous maize gene pairs and their *Tripsacum* homologs.

## MATERIALS AND METHODS

### Plant Material and Tissue Collection

Two mature *Tripsacum* individuals were used for this study. The first individual was from the *T. dactyloides* cultivar Pete (PI 421612), which is a composite of 70 accessions native to Oklahoma and Kansas. Pete was developed through open pollination and combine harvesting over three generations. The rootstock for the Pete individual was acquired from the Tallgrass Prairie Center in Cedar Falls, IA, in 2010 and given the internal identifier IA_Pete_1, passport T0077. The second individual used for this study was a *T. floridanum* field collection obtained near Navy Wells, FL, in 2005. It was given the internal identifier FL_05_15_1, passport T0008. These were mature potted plants that have been clonally propagated in the Cornell University greenhouses set at 22 to 24°C. During this period, they have continually flowered and produced vegetative biomass. The *T. floridanum* tissue types collected for RNA extraction were mature roots, crown tissue (hardened green tissue located at the very base of the leaf bundles, just above the proaxis), whole mature leaves, and an inflorescence from three stages: presilking, postsilking and preanthesis, and postanthesis. The same tissue types were collected for *T. dactyloides* except the postsilking and preanthesis and postanthesis inflorescences. The *T. dactyloides* individual only had one inflorescence at the time of tissue collection. Tissue was harvested in the greenhouse between 2:00 and 3:00 PM and frozen immediately in liquid nitrogen.

### RNA Extraction, Library Preparation, and Sequencing

*Tripsacum* tissue was ground in liquid nitrogen using a mortar and pestle. The RNA was extracted using the Direct-zol RNA miniprep kit with DNase I digestion (Zymo Research). The *Tripsacum* RNA samples and one maize B73 RNA sample were used for library construction. The B73 RNA originated from mature adult leaf grown in a field in Aurora, NY, collected during the day. Strand-specific cDNA libraries were constructed using the SENSE mRNA-Seq library prep kit (Lexogen, Inc.) for Illumina V2. The RTL Reverse Transcription and Ligation Mix was used in the amounts suggested by Lexogen to achieve a mean insert size of 443 bp. Libraries were quantified using the KAPA library quantification kit (KAPA Biosystems) for Illumina platforms. All libraries were multiplexed, and the pool was run on two separate lanes of the NextSeq 500 to generate 2×150 bp reads.

### Read Processing for Quality Control

The following quality control processing steps were performed on the raw reads using Trimmomatic (version 0.32). Adapters listed in Supplemental Table S1 were removed with eight maximum allowed seed mismatches, a palindrome clip threshold of 30, a simple clip threshold of 11, and a minimum adaptor length of one. The forward and reverse reads in the same fragment were kept if they were exact reverse complements of each other. Reads were trimmed if the average Phred score in a five nucleotide window sliding from the 5′ end to the 3′ end fell below 20. Bases were trimmed from the beginning and end of a read if their Phred score was <35 or <20, respectively. All reads shorter than 36 bp were removed from the dataset. To remove contaminating rRNA, these quality-trimmed reads were aligned with Bowtie1 against the MIPS

repetitive sequence element database (Nussbaumer et al., 2013). All reads that aligned to repeats were removed from the dataset. A final effort to remove contaminating Illumina adaptor sequences involved aligning Illumina adaptor sequence queries (Supplemental Table S1) to a BLAST database of cleaned reads. All reads that aligned to adaptor sequences were removed from the dataset (~1400 per library). Finally, to remove inaccurate bases from nonspecific binding of the Lexogen primers during library construction, Trimmomatic clipped 5 bp and 3 bp from the beginning of the forward and reverse read, respectively.

## De novo Assembly and Filtering of the Transcriptomes

After cleaning the reads, all tissue-specific libraries were combined within each species for de novo transcriptome assembly. Trinity (version 2.1.1) assembled the transcriptomes using a minimum k-mer coverage of two. To determine if short transcripts should be removed from the transcriptome assemblies, *Tripsacum* transcripts were aligned to maize B73 protein sequences (RefGen_v3) using BLASTx and an e-value cut-off of $1 \times 10^{-20}$. A Trinity script calculated the percentage length of each maize protein covered by the best *Tripsacum* BLASTx hit. The percentage length of maize homologs covered by transcripts $\geq$500 bp was compared with the percentage length of maize homologs covered by transcripts <500 bp. Transcripts <500 bp were removed from both assemblies because they were mostly fragmented and did not cover the full length of the nearest B73 maize homolog (Supplemental Fig. S1).

Blobtools scripts (Kumar et al., 2013) were used to remove transcripts from the assembly that originated from microbial or fungal species present in the environment and microbiome. Transcript sequences from both species were used as a queries against the BLAST nt database to determine the species of the best BLAST hit. An e-value cutoff of $1 \times 10^{-5}$ was used for the BLAST alignments. All transcripts that did not have a best BLAST hit in Streptophyta, which is the clade of land plants and green algae, were removed from the assembly. Blobtools scripts were also used to map cleaned reads back to the transcriptome assembly of their appropriate species using Bowtie2 (Langmead and Salzberg, 2012) with the '–very-fast-local' parameter. The blobology tool was used to plot each transcript by expression, GC content, and phylum.

## Evaluating the Quality of the Transcriptome Assemblies

To measure how many of the input sequence reads Trinity used in its original assemblies before filtering out low quality and contaminating transcripts, cleaned reads were mapped back to the transcriptome assemblies using Bowtie2. Bowtie2 was used with a maximum insert size of 1000 bp and the '–very-sensitive' parameter.

The L50 statistic was calculated for transcripts at different expression levels. To do this, it was first necessary to quantify transcript abundance within each library. Reads from each tissue-specific library were mapped back to the filtered transcriptomes using the Trinity script 'align_and_estimate_abundance.pl' that calls Bowtie2 and eXpress. A maximum insertion size of 1000 bp was used. For each species, transcripts were divided into 10 different expression bins based on the TMM-normalized transcripts per million statistic reported by eXpress. There were an equal number of transcripts in each expression bin. The L50 statistic was calculated for the transcripts within each expression bin. The definition of L50 is that half of all the assembled bases in each bin exist in transcripts at least as long as the L50 length.

The transcript assemblies were assessed with BUSCO version 2 (Simão et al., 2015). All assembled transcripts within each *Tripsacum* species were submitted as queries against the core eukaryotic gene set for flowering plants (embryophyta_odb9). Duplicate BUSCO hits were removed if they were homologs for the same maize gene. If they had no maize homolog, duplicate BUSCO hits were removed if they were grouped within the same Trinity gene cluster.

## Finding Maize Homologs of *Tripsacum* Transcripts

*Tripsacum dactyloides* and *T. floridanum* transcripts were aligned to the B73 RefGen_v4 genome using the Program to Assemble Spliced Alignments (PASA) (Haas et al., 2003). The PASA tool was run using gmap as the aligner with the following requirements for a valid alignment: a maximum intron length of 20,000 bp, a minimum average percentage identity of 80%, a minimum aligned length of 70%, and no base pairs flanking the splice junctions were required to match perfectly. The PASA tool reported only one valid alignment per transcript.

A *Tripsacum* transcript was considered a fully assembled homolog of a maize gene if the following were true: (i) part of the transcript alignment fell within the boundaries of the maize gene and aligned to the same genomic strand as the maize gene, (ii) the best BLASTx hit for the *Tripsacum* transcript was a protein for that maize gene (RefGen_v3), and (iii) the *Tripsacum* transcript alignment covered at least 95% of the length of the best BLASTx maize protein hit (RefGen_v3). A single maize transcript from each maize gene homolog was considered the best homolog for a *Tripsacum* transcript if the following were true: (i) the maize RefGen_v4 transcript aligned to the same maize RefGen_v3 protein as the fully assembled *Tripsacum* transcript, (ii) the maize RefGen_v4 transcript and maize RefGen_v3 protein were expressed from the same gene according to the gene ID history between version 3 and version 4 of the maize genome (using the file available at the time: maize.v3TOv4.geneIDhistory.txt), and (iii) the RefGen_v4 transcript had the longest alignment length to the RefGen_v3 protein out of any other transcript from the RefGen_v4 gene. The RefGen_v3 annotations were used because the RefGen_v4 annotations were under revision at the time. The transcript homolog pairs that were between 500 and 10,000 bp long were analyzed for GC content using in-house scripts available on Bitbucket (https://bitbucket.org/bucklerlab/tripsacumtranscriptomes/src/master/).

### Characterizing *Tripsacum* Transcripts Without Maize Homologs

We sought to identify transcripts that were present in *Tripsacum* and absent from maize. *Tripsacum* transcripts that did not align to the B73 RefGen_v4 genome using PASA were aligned to other sequence databases to determine whether they had maize homologs. First, these transcripts were aligned against maize B73 RefGen_v3 cDNA using BLASTn. A discontiguous megablast parameter and an e-value threshold of $1 \times 10^{-20}$ were used. Next, these transcripts that did not align to B73 RefGen_v4 were aligned against the B104 (http://ftp.maizegdb.org/MaizeGDB/FTP/B104/Pseudomolecule_Assembly/), CML247 (http://ftp.maizegdb.org/MaizeGDB/FTP/CML247/), EP1 (https://maizegdb.org/genome/genome_assembly/Zm-EP1-REFERENCE-TUM-1.0), F7 (https://maizegdb.org/genome/genome_assembly/Zm-F7-REFERENCE-TUM-1.0), Mo17 (https://www.maizegdb.org/genome/genome_assembly/Zm-Mo17-REFERENCE-YAN-1.0), PH207 (Hirsch et al., 2016), and W22 (ftp://ftp.ncbi.nlm.nih.gov/genomes/genbank/plant/Zea_mays/latest_assembly_versions/GCA_001644905.2_Zm-W22-REFERENCE-NRGENE-2.0) inbred genomes using PASA. The PASA tool was run using gmap as the aligner, a maximum intron length of 20,000 bp, a minimum average percentage identity of 80%, a minimum aligned length of 70%, and no base pairs flanking the splice junctions were required to match perfectly. Finally, the *Tripsacum* transcripts that did not align to B73 RefGen_v4 were aligned to the BLAST nt database using Blobtools scripts. *Tripsacum* transcripts that did not align to any of the eight maize inbred reference genomes, B73 cDNA, or other *Zea mays* genotype in the BLAST nt database were considered not to have a maize homolog.

The expression values of transcripts were obtained from the previous eXpress run. Transcripts with an FPKM > 1 (FPKM, fragments per kilobase of transcript per million mapped reads) were used as a BLASTx query against the NCBI nonredundant protein database. The reading frame with the best BLAST hit was considered the open reading frame. Reverse open reading frames were not used because the transcripts were assembled in a strand-specific manner. The EMBOSS transeq (Rice et al., 2000) software was used to get translations in the open reading frame. Conserved protein domains within these translations were identified using the HMMER (version 3.1b2) hmmscan command. The protein sequences were used as a query against the Pfam A database with an inclusion threshold of 0.01 (Eddy, 2009; Finn et al., 2016).

### Comparing B73 and *Tripsacum* Expression

The following quality control processing steps were performed on the raw B73 reads using Trimmomatic (version 0.36). Adapters listed in Supplemental Table S1 were removed with eight maximum allowed seed mismatches, a palindrome clip threshold of 30, a simple clip threshold of 11, and a minimum adaptor length of one. Both reads were kept if they were exact reverse complements of each other. All reads <36 bp were removed from the dataset. To remove inaccurate bases from nonspecific binding of the Lexogen primers during library construction, Trimmomatic clipped 9 and 6 bp from the beginning of the forward and reverse read, respectively. Reads were trimmed if the average Phred score in a five-base sliding window fell below 20. Bases were trimmed from the end of a read if their Phred score was <20.

Mature adult leaf libraries from B73, *T. dactyloides*, and *T. floridanum* were aligned to the B73 genome (RefGen_v4) using the STAR aligner (Dobin et al. [2012], version 2.5.2b). Read alignments were considered valid if they were unique and <6% of the read was mismatches. The RefGen_v4.34 annotations were during the STAR alignments. Each gene in the gene transfer format (GTF) file had its 3′ boundary extended by 500 bp. The '–sjdbOverhang' parameter was used with a value of 89. Cufflinks2 (version 2.2.1) was then used to estimate transcript abundance. The intron size was limited to 20 to 80,000 bp. Gene FPKM values from Cufflinks2 were used to compare *Tripsacum* and maize expression. Genes with FPKM values lower than 0.01 in either species were not used for the comparison.

Three gene expression comparisons were made for homologous genes: (i) genes in the maize proteome (Walley et al., 2016) and their *Tripsacum* homologs vs. genes that were not detected in the maize proteome and their *Tripsacum* homologs, (ii) maize subgenome 1 genes and their *Tripsacum* homologs vs. maize subgenome 2 genes and their *Tripsacum* homologs, and (iii) genes in core functional pathways vs. genes functioning in stress pathways. Genes in core functional pathways were defined as genes having gene ontology (GO) terms associated with photosynthesis, electron transport, cytochrome C, DNA repair, glycolysis, meiosis, mitosis, recombination, and starch metabolism. Genes in stress pathways were defined as genes having GO terms associated with cold, defense, disease, drought, pathogen, resistance, and stress. The gene lists for these categories and scripts for comparing expression are publicly available on Bitbucket (https://bitbucket.org/bucklerlab/tripsacum-transcriptomes/src/master/). Genes with FPKM values >0.01 were $\log_{10}$–transformed, and a two-sample *t*-test with the Welch approximation was used to test whether gene sets were differently expressed.

Expression differences between genes in maize homeologous pairs were calculated for the high-confidence homeologous pairs identified by Schnable et al. (2011). Their RefGen_v4 names were identified using the maizeGDB conversion file v3_v4_xref.txt.

## RESULTS

### De novo Transcriptome Assembly

Transcriptomes were assembled from root, leaf, crown, and inflorescence tissue from two species, *T. dactyloides* and *T. floridanum,* using Trinity (Grabherr et al., 2011). Transcripts <500 bp were filtered out of the assemblies because they were mostly fragmented and did not cover the full length of the nearest B73 maize homolog

Table 1. De novo transcriptome assembly statistics for *Tripsacum dactyloides* and *T. floridanum*.

| Assembly statistic | *T. dactyloides* | *T. floridanum* |
|---|---|---|
| Amount of high-quality sequence used for assembly | 46 Gbp | 79 Gbp |
| Reads mapped to unfiltered transcriptome assembly | 93% | 93% |
| Transcript L50 | 1442 bp | 1452 bp |
| Median transcript length | 981 bp | 956 bp |
| Mean transcript length | 1230 bp | 1233 bp |
| Number of trinity-defined genes | 50,411 | 64,422 |
| Number of trinity-defined transcripts | 131,952 | 155,705 |
| Completely assembled BUSCO genes | 83.8% | 85.5% |
| Total No. of cleaned reads per sample | 154,999,184 | 261,829,615 |
| No. of cleaned reads for root library | 40,094,172 | 44,171,419 |
| No. of cleaned reads for crown library | 47,851,949 | 41,257,696 |
| No. of cleaned reads for leaf library | 34,149,344 | 49,970,164 |
| No. of cleaned reads for presilking inflorescence library | 32,903,719 | 39,739,865 |
| No. of cleaned reads for postsilking or preanthesis inflorescence library | – | 40,940,770 |
| No. of cleaned reads for postanthesis library | – | 45,749,701 |

(Supplemental Fig. S1). Contaminating transcripts from organisms such as fungi and insects in the environment or microbiome were removed using Blobtools scripts (Supplemental Fig. S2) (Kumar et al., 2013).

In the final transcriptome assemblies, Trinity grouped the transcripts into 50,411 *T. dactyloides* genes and 64,422 *T. floridanum* genes (Table 1), which is considerably more than the 39,656 genes in the filtered RefGen_v3 maize gene set (Law et al., 2015). Trinity may have overestimated gene number if it misidentified allelic transcripts as transcripts originating from different genes. The sequenced individuals are highly heterozygous, so allelic transcripts were probably assembled for many genes. We do not attempt to collapse allelic transcripts at the gene level because our ultimate purpose was to identify homologous pairs at the transcript level in maize and *Tripsacum*. The final assemblies contain 131,952 *T. dactyloides* transcripts and 155,705 *T. floridanum* transcripts, respectively (Table 1).

## The Transcriptome Assemblies are High Quality

Ideally, the average transcript length of a de novo transcriptome assembly should be similar to the average transcript length in the most closely related species. The mean transcript lengths were 1230 bp for *T. dactyloides* and 1233 bp for *T. floridanum*, while the mean transcript length is 1541 bp in maize RefGen_v3 annotations (Law et al., 2015). The L50 was 1442 and 1452 bp for the *T. dactyloides* and *T. floridanum* assemblies, respectively. The L50 statistic defines the transcript length for which half of all assembled bases exist in transcripts longer than the L50 length. More highly expressed transcripts have longer L50 lengths, probably because higher read coverage depth enables a more complete assembly (Supplemental Fig. S3).

The Benchmarking Universal Single-Copy Ortholog (BUSCO) gene set for land plants was used to assess the completeness of the two assemblies (Simão et al., 2015). Of all plant BUSCO genes, 83.8 and 85.5% were completely assembled in *T. dactyloides* and *T. floridanum*, respectively (Supplemental Fig. S4). Thus, a majority of genes that are highly conserved across land plants are represented in the two assemblies.

## Identification of *Tripsacum* and Maize Homologs

*Tripsacum dactyloides* and *T. floridanum* transcripts were aligned to the B73 RefGen_v4 genome using PASA (Haas et al., 2003) to identify their nearest maize homologs. Alignments were required to have at least 80% identity between maize and *Tripsacum* and at least 70% of the *Tripsacum* transcript length aligning to the genome. The average percentage identity across all valid transcript alignments is 91.98% in *T. dactyloides* and 91.49% in *T. floridanum*. There are 57,080 *T. dactyloides* transcripts aligned to 18,327 maize genes and 58,973 *T. floridanum* transcripts aligned to 18,903 maize genes (Supplemental Data 1,2). There are 6688 maize genes with transcripts from more than one *T. dactyloides* Trinity gene mapping to them. There are 16,716 *T. dactyloides* Trinity genes aligning to these 6688 maize genes. Likewise, 8046 maize genes have transcripts from more than one *T. floridanum* Trinity gene mapping to them. There are 21,058 *T. floridanum* Trinity genes aligning to these 8046 maize genes. Multiple Trinity-defined Tripsacum genes having the same nearest maize homolog is evidence for Trinity mischaracterizing alleles as separate genes.

Transcripts that did not align to the B73 genome or B73 cDNA were aligned against the B104, CML247, EP1, F7, Mo17, PH207, and W22 maize inbred genomes using PASA. *Tripsacum* transcripts that did not align to any of the eight maize inbred reference genomes, B73 cDNA, or other *Zea mays* genotype in the BLAST nt database were considered not to have a maize homolog. There are 1671 *T. dactyloides* transcripts and 2220 *T. floridanum* transcripts without a maize homolog that had an expression level greater than one FPKM in at least one tissue, which are listed in Supplemental Data 3,4, along with their best BLAST hits from the NCBI nonredundant protein database and their conserved Pfam domains. These *Tripsacum* transcripts without maize homologs contain a broad variety of conserved protein domains and a diverse set of functions, and many are putative transcription factors.

A smaller subset of *Tripsacum* transcripts with maize homologs were identified as fully assembled, meaning that the transcripts aligned to 95% of the length of its nearest maize protein homolog (RefGen_v3). There are 6982 homologous fully assembled transcript pairs between *T. dactyloides* and maize, and there are 6368 homologous fully assembled transcript pairs between *T. floridanum* and maize.

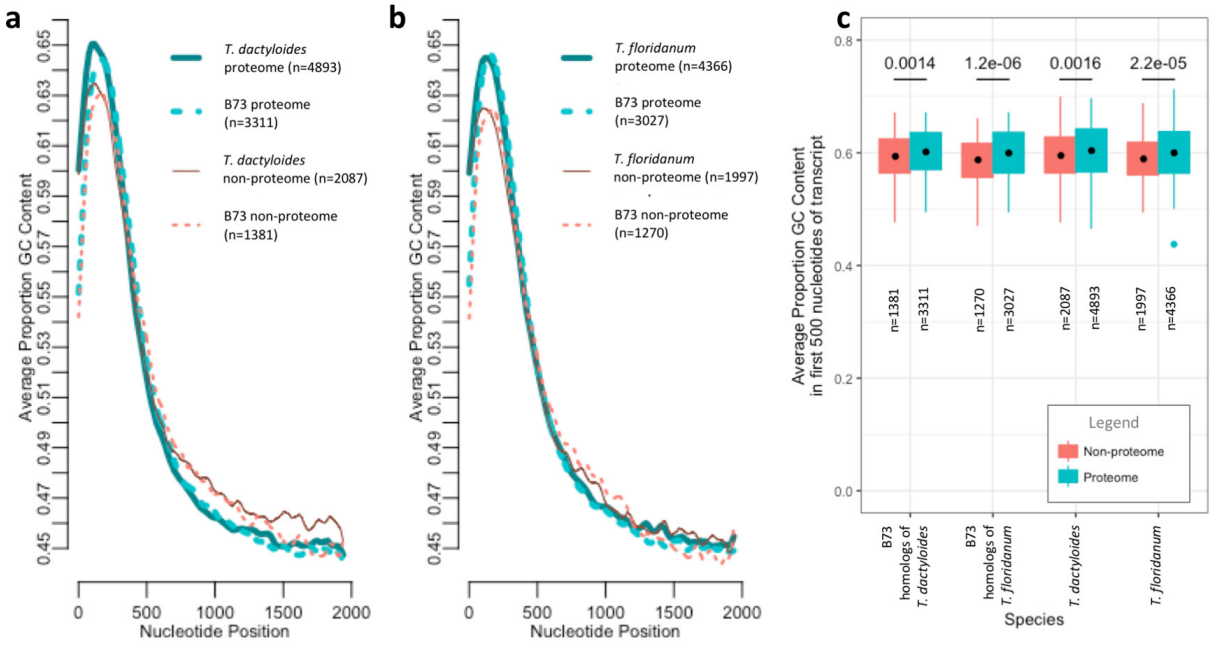


Fig. 1. Average guanine–cytosine (GC) content in fully assembled *Tripsacum* transcripts and their maize homologs. (a,b) LOWESS curves of GC content averaged across transcripts at each nucleotide position. Homologous transcript pairs for maize and (a) *T. dactyloides* or maize and (b) *T. floridanum* were split into two subsets: pairs where the maize homolog is detected in the proteome (Walley et al., 2016) and pairs where the maize homolog is not detected in the proteome. (c) Boxplot of the average proportion GC content in the first 500 nucleotides of transcripts. Values for *P* are shown for the Student's two-sided *t*-test with the Welch approximation. The turquoise point is an outlier, and the black points represent the mean of the average proportion GC content across the first 500 nucleotides.

## Proteome Transcripts have Higher GC Content than Nonproteome Transcripts in Maize and *Tripsacum*

We hypothesized that genes with protein evidence tend to be more resistant to pseudogenization than genes with no detectable protein. Genes without detectable protein products are likely enriched for pseudogenes, which, by definition, do not produce protein. Indeed, a comprehensive proteomics analysis has revealed that genes without detected proteins have a higher proportion of annotated pseudogenes than genes with expressed proteins (Walley et al., 2016). Additionally, maize genes that show syntenic conservation with sorghum are nine times more likely to express protein (Walley et al., 2016).

Genes that are resistant to pseudogenization in maize and *Tripsacum* are predicted to be located near recombination hotspots, and GC content can reflect recombination rate. We tested whether protein-encoding maize transcripts and their *Tripsacum* homologs (hereafter called proteome pairs) have maintained higher GC content than putatively untranslated maize transcripts and their *Tripsacum* homologs (hereafter called nonproteome pairs). The proteome study by Walley et al. (2016) was used to determine whether a maize transcript encoded a protein or not. Average GC content was calculated along the length of maize and fully assembled *Tripsacum* homologous transcripts at all coding and noncoding positions. The GC content peaks at the 5′ end of transcripts and declines toward the 3′ end in maize, *T. dactyloides*, and *T. floridanum* (Fig. 1a,b). Other studies have also observed high GC content at the 5′ end of genes and a declining GC gradient over the length of the gene in monocots and dicots (Serres-Giardi et al., 2012; Wong et al., 2002). Recombination hotspots in monocots and dicots tend to be located at the 5′ end of genes and, to a lesser extent, at the 3′ end of genes (Hellsten et al., 2013; Choi et al., 2013; Singh et al., 2016). Thus, the observed declining GC gradient over transcript length is probably a result of GC-biased gene conversion. Proteome pairs had about a 1 to 2% higher GC content peak at the 5′ end of transcripts than nonproteome pairs (Fig. 1a–c). These results may indicate that the regions of the genome where proteome pairs reside have higher recombination rates than the regions of the genome where nonproteome pairs reside. However, further experiments that measure recombination rates in proteome and nonproteome pairs would be needed to reach this conclusion

## Proteome Genes have Higher, More-Consistent Expression than Nonproteome Genes across Maize and *Tripsacum*

If proteome genes have higher expression than nonproteome genes, then the hypothesis that proteome genes are more resistant to gene loss in maize and *Tripsacum* will be supported because highly expressed genes are more resistant to genome fractionation than lowly expressed genes in maize (Schnable et al., 2011; Zhao et al., 2017). Gene expression levels were estimated using maize and *Tripsacum* RNA-seq reads that mapped uniquely to the maize genome. The percentage of reads that mapped uniquely for each species was 60.81% in *T. dactyloides*, 53.86% in

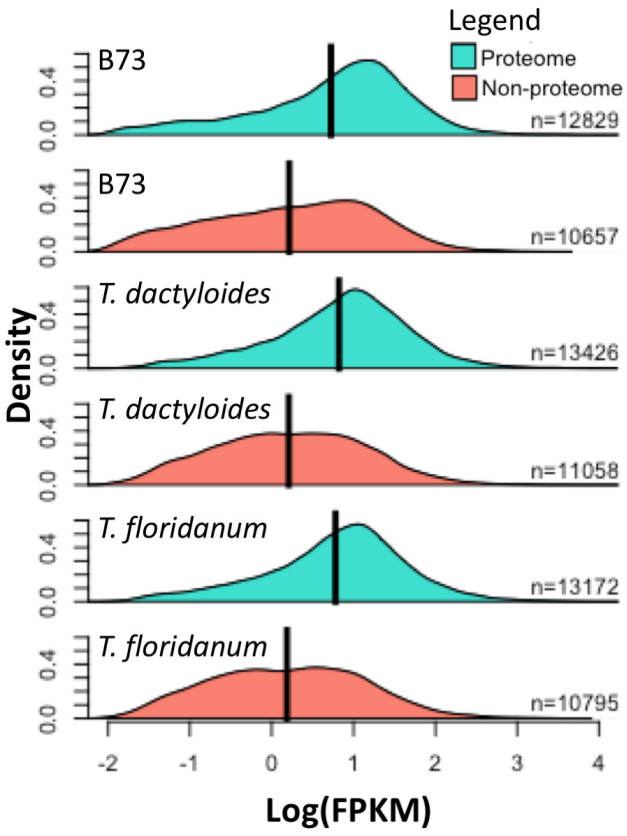


Fig. 2. Probability density plots of gene expression levels in maize inbred B73, *Tripsacum dactyloides*, and *T. floridanum*. The black line is the mean log-transformed expression level. The RNA-seq reads from mature adult leaves were aligned to the B73 genome (RefGen_v4) using Cufflinks2. Maize genes were divided into two subsets: genes translated into the maize proteome (Walley et al., 2016) and genes not detected in the proteome. The log-transformed expression of genes with an FPKM > 0.01 are plotted. FPKM, (fragments per kilobase of transcript per million mapped reads).

*T. floridanum*, and 81.72% in maize. The average alignment mismatch rate was 2.58% in *T. dactyloides*, 2.57% in *T. floridanum*, and 0.19% in maize. Homologous genes with FPKM values >0.01 in both species are plotted in Fig. 2. A two-sample *t*-test with the Welch approximation was performed to test whether protein-encoding maize genes and their *Tripsacum* homologs (proteome pairs) were expressed at different levels than nonproteome maize genes and their *Tripsacum* homologs (nonproteome pairs). Proteome pairs had significantly higher mean log-transformed FPKM values than nonproteome pairs in maize ($t$[22176] = 39.7; $p < 2.2 \times 10^{-16}$), *T. dactyloides* ($t$[22728] = 54.3; $p < 2.2 \times 10^{-16}$), and *T. floridanum* ($t$[22332] = 50.6; $p < 2.2 \times 10^{-16}$) (Fig. 2). The average untransformed maize expression level was 30.0 FPKM for proteome genes and 13.3 FPKM for nonproteome genes. The average untransformed *T. dactyloides* expression level was 36.5 FPKM for homologs of proteome genes and 14.4 FPKM for homologs of nonproteome genes. The average untransformed *T. floridanum* expression level was 39.0 FPKM for homologs of proteome genes and 14.8 FPKM for homologs of nonproteome genes. Thus, proteome maize genes and their *Tripsacum* homologs are more highly expressed than nonproteome maize genes and their *Tripsacum* homologs.

Other gene sets also show similarities in gene expression across all three species. For example, genes with GO terms relating to housekeeping functions, such as photosynthesis, glycolysis, and mitosis, are more highly expressed in nonstress conditions relative to genes with stress-related GO terms in B73 ($t$[3537.1] = 8.0672; $p = 9.759 \times 10^{-16}$), *T. dactyloides* ($t$[3369] = 2.8303; $p$ = 0.004678), and *T. floridanum* samples ($t$[3383.6] = 4.2366; $p = 2.33 \times 10^{-5}$) (Supplemental Fig. S5). Surprisingly, maize subgenome 1 genes as a group do not show significantly higher expression than maize subgenome 2 genes as a group in B73 ($t$[17074] = 0.7253; $p$ = 0.4683), *T. dactyloides* homologs ($t$[18111] = 0.39181; $p$ = 0.6952), or *T. floridanum* homologs ($t$[17811] = 0.8799; $p$ = 0.3789) (Supplemental Fig. S6). Over 13,000 maize subgenome 1 genes and *Tripsacum* homologs and over 8000 maize subgenome 2 genes and *Tripsacum* homologs were included in the expression comparison. Schnable et al. (2011) found that subgenome 1 genes tended to be more highly expressed than their subgenome 2 homeolog in 1750 high-confidence homeologous pairs in eight tissues. *Tripsacum* homologs were identified for 1625 high-confidence maize homeolog pairs from Schnable et al. (2011), and we were able to replicate these results in our dataset; the *Tripsacum* homolog of the maize subgenome 1 gene also tended to exhibit expression dominance more often than the *Tripsacum* homolog of the maize subgenome 2 gene (Supplemental Fig. S7). However, expanding the analysis beyond high-confidence homeologous pairs to all subgenome 1 and subgenome 2 genes reveals that maize subgenome 1 genes as a whole are not more highly expressed than maize subgenome 2 genes as a whole and neither are their *Tripsacum* homologs in the leaf samples measured in this study (Supplemental Fig. S6).

We hypothesized that proteome pairs had more consistent gene expression levels between maize and *Tripsacum* than nonproteome pairs. Simple linear regressions were performed to predict *T. dactyloides* (Fig. 3a–c) and *T. floridanum* (Fig. 3d–f) gene expression based on maize gene expression. Expectedly, maize gene expression showed higher correlation among proteome genes ($R^2$ = 0.42 for *T. dactyloides*, $R^2$ = 0.49 for *T. floridanum*) than nonproteome genes ($R^2$ = 0.29 for *T. dactyloides*, $R^2$ = 0.37 for *T. floridanum*). Simple linear regressions were also performed to predict *T. dactyloides* gene expression based on *T. floridanum* gene expression (Fig. 3 g–i). *Tripsacum dactyloides* gene expression showed high correlation with *T. floridanum* both among proteome genes ($R^2$ = 0.78) and nonproteome genes ($R^2$ = 0.72). This indicates that proteome pairs have maintained more consistent gene expression across genera; meanwhile, the expression levels of nonproteome homologs are relatively consistent within the *Tripsacum* genus but have become more dissimilar across the *Tripsacum* and maize genera.

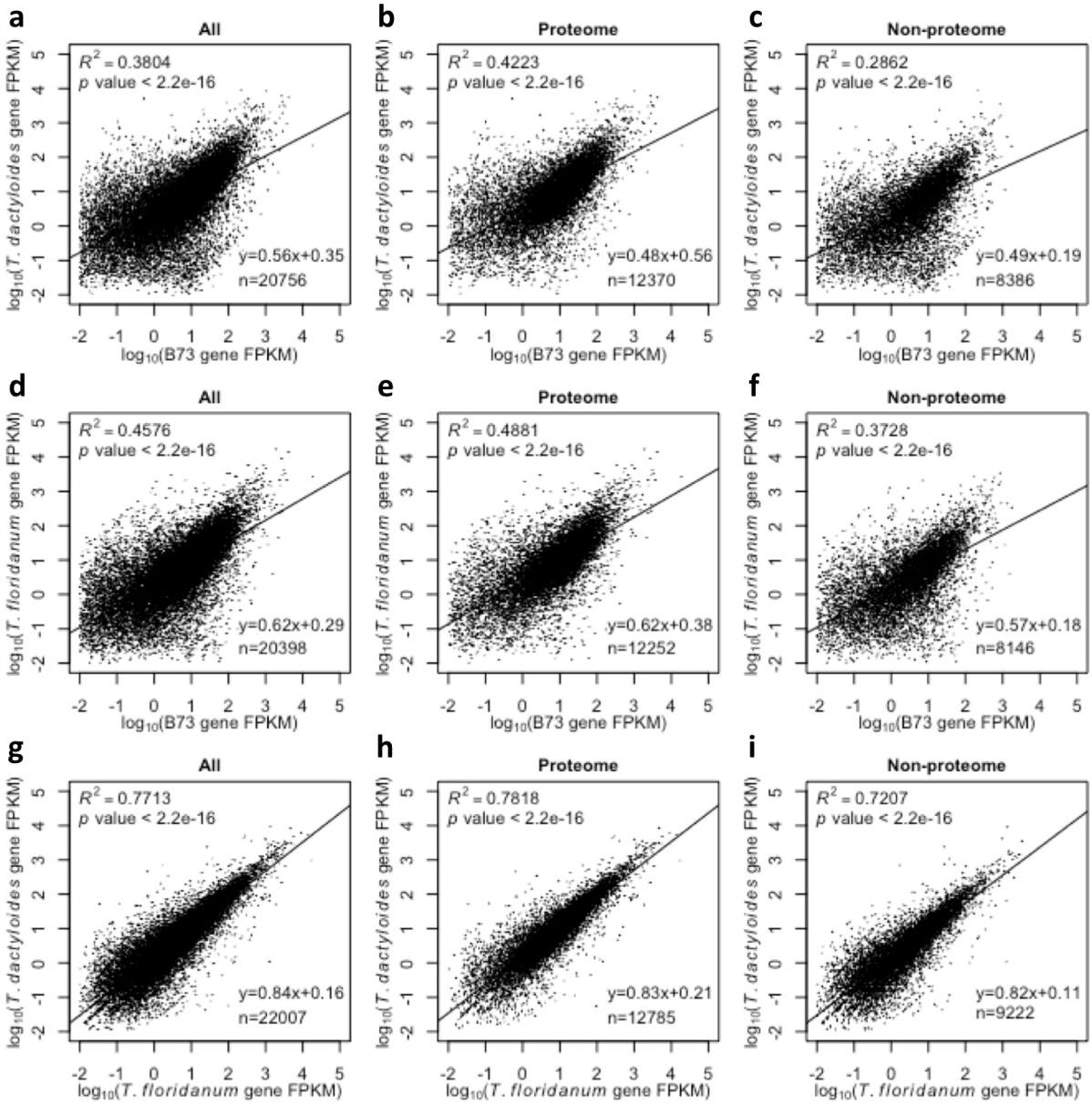


Fig. 3. Correlation of gene expression levels in maize inbred B73, *Tripsacum dactyloides*, and *T. floridanum*. The RNA-seq reads from mature adult leaves were aligned to the B73 genome (RefGen_v4) using Cufflinks2. Maize genes were divided into two subsets: genes translated into the maize proteome (Walley et al., 2016) and genes not detected in the proteome. The log-transformed expression of genes with an FPKM > 0.01 were plotted in (a–c) B73 vs. *T. dactyloides*, (d–f) B73 vs. *T. floridanum*, and (g–i) *T. floridanum* vs. *T. dactyloides*. Adjusted $R^2$ values are shown.

Like genes in the maize proteome, genes with essential functions also maintain consistent gene expression across maize and *Tripsacum*. Core housekeeping gene expression showed higher correlation across species ($R^2$ = 0.50 for B73 vs. *T. dactyloides*, $R^2$ = 0.55 for B73 vs. *T. floridanum*, $R^2$ = 0.83 for *T. dactyloides* vs. *T. floridanum*) than stress-related genes ($R^2$ = 0.34 for B73 vs. *T. dactyloides*, $R^2$ = 0.40 for B73 vs. *T. floridanum*, $R^2$ = 0.75 for *T. dactyloides* vs. *T. floridanum*) (Supplemental Fig. S8). In contrast, maize subgenome 1 genes and their *Tripsacum* homologs do not have higher correlation across species ($R^2$ = 0.38 for B73 vs. *T. dactyloides*, $R^2$ = 0.45 for B73 vs. *T. floridanum*, $R^2$ = 0.77 for *T. dactyloides* vs. *T. floridanum*) than maize subgenome 2 genes and their *Tripsacum* homologs ($R^2$ = 0.39 for B73 vs. *T. dactyloides*, $R^2$ = 0.47 for B73 vs. *T. floridanum*, $R^2$ = 0.79 for *T. dactyloides* vs. *T. floridanum*) (Supplemental Fig. S9).

## The Same Gene Tends to Have Higher Expression in Reciprocally Retained Maize Homeolog Pairs and Putative *Tripsacum* Homeolog Pairs

If genome fractionation is occurring in a similar fashion in maize and *Tripsacum*, then the same member of a homeolog pair will exhibit higher expression than the other member in both clades. There are 1625 homeolog pairs identified by Schnable et al. (2011) that have homologs in our *T. dactyloides* and *T. floridanum* transcriptomes. Each maize homeolog pair consists of a homeolog from the maize subgenome 1 and a homeolog from the maize subgenome 2. The

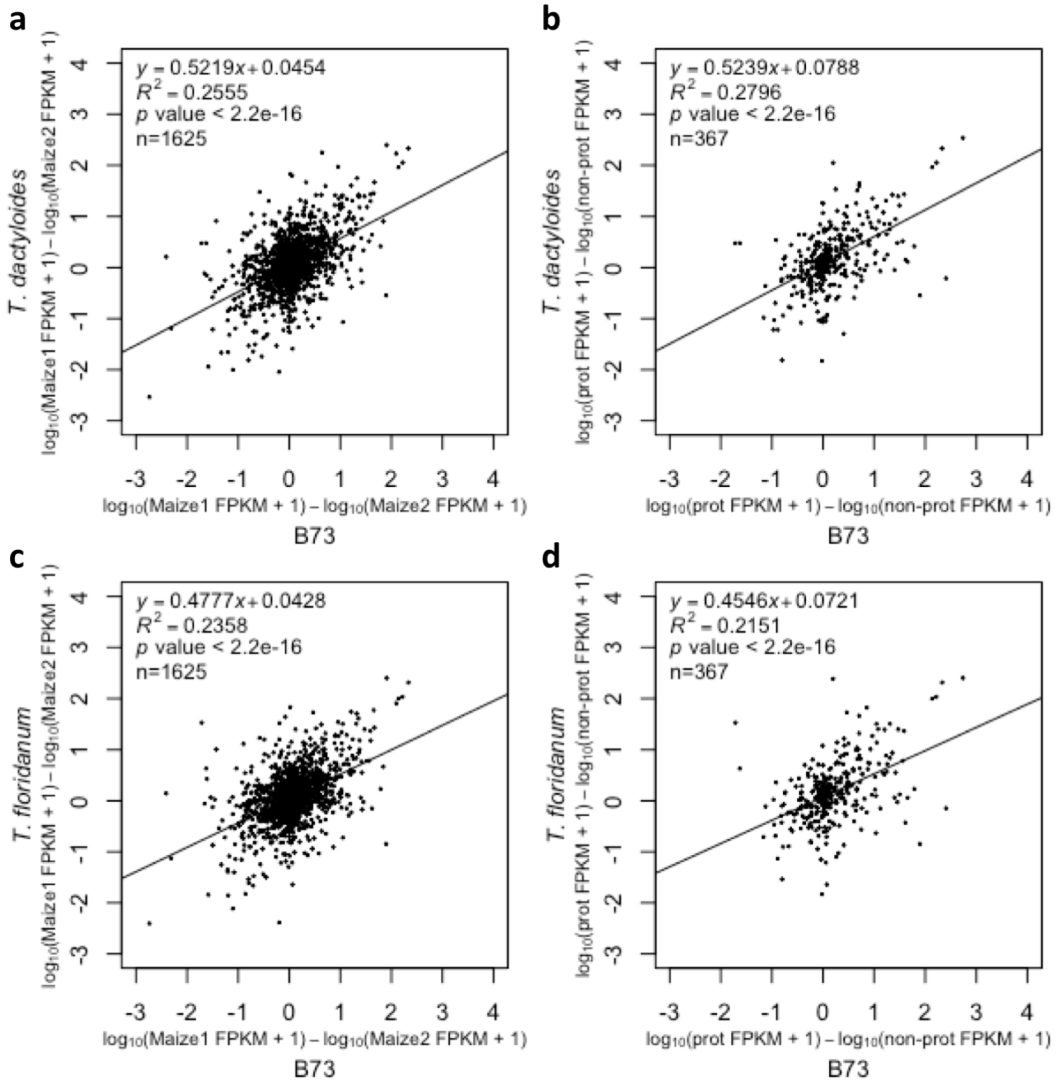


Fig. 4. Correlation of homeolog expression differences in (a,b) maize inbred B73 and *Tripsacum dactyloides* and (c,d) maize inbred B73 and *T. floridanum*. (a,c) Difference in log-transformed FPKM (fragments per kilobase of transcript per million mapped reads) values between maize subgenome 1 homeolog and maize subgenome 2 homeolog is plotted for all 1625 homeolog pairs. (b,d) Difference in log-transformed FPKM values between the homeolog detected in the proteome and the homeolog not detected in the proteome is plotted for 367 homeolog pairs. A unit of 1 FPKM was added to all gene expression values to aid in log transformation of unexpressed genes. Adjusted $R^2$ values are shown.

*Tripsacum* genome has not been assembled, thus syntenic paralogs cannot be confidently identified for *Tripsacum* subgenome 1 or *Tripsacum* subgenome 2. Consequently, the *Tripsacum* homologs that have the highest sequence similarity to maize subgenome 1–subgenome 2 homeologous pairs will be called putative *Tripsacum* homeologs.

The expression difference between maize homeologs was significantly and positively correlated with the expression difference between putative *Tripsacum* homeologs (Fig. 4a,c). Of the 367 maize homeolog pairs that had one proteome gene and one gene that was not detected in the maize proteome, the homeolog expression difference was significantly and positively correlated with the expression difference between putative *Tripsacum* homeologs (Fig. 4b,d). These data may suggest that the same homeolog tends to show expression dominance in maize and *Tripsacum*, although a sequenced *Tripsacum* genome is needed to accurately identify *Tripsacum* homeologous pairs.

Expression levels often differ between members of homeologous gene pairs. In soybean, about half of homeologous gene pairs exhibited transcriptional divergence between members in at least one tissue (Roulin et al., 2013). In a natural cotton allopolyploid (*Gossypium L.*), about 40% of 49 homeologous gene pairs showed transcriptional bias toward one homeolog or the other (Chaudhary et al., 2009). In maize, 98% of 3228 homeologous gene pairs exhibited two-fold transcriptional divergence between members across multiple tissues (Pophaly and Tellier, 2015). The more highly expressed maize homeolog had significantly lower ΠN/ΠS ratios, and thus were under stronger purifying selection, than the more lowly expressed homeolog (Pophaly and Tellier, 2015). In this study, we do not test for differential expression or ΠN/ΠS ratios of genes within homeologous pairs, but we do see that the polarity in maize homeolog expression is conserved in their *Tripsacum* homologs.

## DISCUSSION

It is unknown whether the gene complements of maize and *Tripsacum* are reducing in size in a similar fashion following their shared ancient tetraploidy event. The 15,160 maize genes with protein evidence are promising candidates for genes resistant to pseudogenization in both clades. Maize genes that do not produce detectable protein may be enriched for pseudogenes, which by definition do not express protein. Maize genes known to express protein tend to show syntenic conservation with sorghum (Walley et al., 2016), and syntenic genes tend to have more highly conserved expression patterns and less presence–absence variation (Schnable, 2015). With this study, we show that maize proteome genes have higher GC content and higher expression than nonproteome maize genes (Fig. 1,2). Furthermore, *Tripsacum* homologs of maize proteome genes also have higher GC content and gene expression than *Tripsacum* homologs of maize nonproteome genes (Fig. 1,2). Expression patterns are more consistent across maize and *Tripsacum* for proteome pairs than nonproteome pairs (Fig. 3). The same homeolog or putative homeolog tends to show expression dominance in both maize and *Tripsacum* (Fig. 4). Together, these data highlight similarities in maize and *Tripsacum* gene evolution after whole-genome duplication.

In this study, GC content served as an indirect measure of recombination rate because exonic and intronic GC content is positively correlated with recombination rate resulting from GC-biased gene conversion (Muyle et al., 2011). The GC content in exonic positions peaked near the 5′ end of the gene and steeply declined toward the 3′ end. Because recombination hotspots in monocots and dicots tend to be located at the 5′ end of genes (Hellsten et al., 2013; Choi et al., 2013; Singh et al., 2016), this decline in GC content likely is due to GC-biased gene conversion. The higher the heterozygosity, the stronger the effect of GC-biased gene conversion (Muyle et al., 2011). Thus, GC-biased gene conversion is predicted to have a large effect on GC content because *Tripsacum* and maize populations are highly heterozygous. Although proteome pairs have higher GC content than nonproteome pairs, recombination breakpoints need to be measured to determine whether proteome pairs truly reside in genomic regions with higher recombination rates than nonproteome pairs.

One potential limitation of this study is that a single biological leaf replicate was used for the maize, *T. dactyloides,* and *T. floridanum* expression analyses. Single replicates were taken from root, leaf, crown, and inflorescence tissue with the purpose of sampling transcript diversity to assemble the *Tripsacum* transcriptomes. When the leaf RNA-seq data was later used for interspecific comparisons of expression, the same patterns in gene expression between maize and *T. dactyloides* are also evident when comparing maize and *T. floridanum*. *Tripsacum dactyloides* and *T. floridanum* behave as independent replicates for the *Tripsacum* genus. Both replicates indicate that maize proteome gene expression is more highly correlated with *Tripsacum* homolog expression than maize nonproteome gene expression. Furthermore, each homeolog pair or putative homeolog pair represents a separate opportunity for expression divergence to occur. The same homeolog tended to have higher expression in each of the three species.

While the exact nature of maize genes lacking protein evidence needs further study, this gene set may be enriched for pseudogenes or genes on the evolutionary path toward pseudogenization. Pseudogenes can form after a polyploidy event when there is a lack of selective pressure to maintain two copies, and one homeolog loses the ability to produce a functional protein (Xiao et al., 2016). Although the two homeologs have similar gene sequences after whole-genome duplication, they differ in genomic location as large-scale chromosomal rearrangements occur. If one homeolog has a lower recombination rate, it is more likely to accumulate deleterious mutations. Mutations or small intraexon deletions (Woodhouse et al., 2010) that disrupt the open reading frame can cause frame shifts or premature termination codons and may transform a gene into a pseudogene. Several studies have shown that pseudogenes can be actively transcribed. Messenger RNA was detected for 12% of rice pseudogenes, which have significantly lower expression levels than their functional paralogs (Thibaud-Nissen et al., 2009). Out of 129 lineage-specific pseudogenization events in four Poaceae species, 61 pseudogenes are actively transcribed (Zhao et al., 2015). Zou et al. (2009) found that 17 and 23% of pseudogenes are expressed in rice and *Arabidopsis*, respectively, and they have similar or lower expression than functional genes. Pseudogenes may be misannotated as functional genes using ab initio gene finding software, which can alter gene structure to avoid frame shift mutations or premature stop codons (Thibaud-Nissen et al., 2009).

Despite a thorough sampling of 33 tissues, some proteins may be missing from the Walley et al. (2016) proteome if they accumulate in stress conditions or at certain times in the diurnal cycle that were not sampled. Some proteins may not have been detected because of limited sensitivity of the mass spectrometer. Thus, some genes we have been referring to as nonproteome genes may not be pseudogenes because they do in fact produce functional proteins. More experiments are needed to determine whether these genes without detectable protein are truly undergoing pseudogenization.

The expression analysis between 1625 high-confidence homeolog pairs identified by Schnable et al. (2011) in maize and their *Tripsacum* homologs reveals that the same member of the pair tends to have expression dominance in both maize and *Tripsacum* (Fig. 4). The proteome member does not always have higher expression than the nonproteome member. Similarly, the maize 1 homeolog or its *Tripsacum* homolog does not always have higher expression than the maize 2 homeolog or its *Tripsacum* homolog. Likewise, it may be expression dominance itself that determines whether a homeolog is more likely to resist fractionation.

Natural selection is shaping the gene sets of maize and *Tripsacum* similarly after tetraploidy. Maize breeding programs could benefit from knowing which genes encode protein in maize and are highly and consistently expressed across maize and *Tripsacum*. Markers in these genes could be

appropriately weighted for genomic selection because they are more likely to control plant phenotype. This study represents one application out of many possible applications for how the de novo transcriptome assemblies presented here can be used to provide insight into maize genome evolution.

## Supplemental Material

Supplemental material includes supplemental figures and tables and is available online.

## Data Deposition

The transcriptome assemblies reported in this paper are available at https://doi.org/10.7946/P23P8K. The accession numbers for the transcriptome assemblies reported in this paper will be publicly available on publication. Scripts used in this study are publicly available in the following Bitbucket repository: https://bitbucket.org/bucklerlab/tripsacumtranscriptomes/src/master/

## Pre-print

A pre-print is available on biorXiv with manuscript ID BIORXIV/2018/267682.

## Author Contributions

C.M.G. and E.S.B. designed research. C.M.G. performed research. C.M.G., K.A.K., and E.S.B. analyzed data. C.M.G. wrote the paper.

## Conflict of Interest Disclosure

The authors declare no conflicts of interest.

## ACKNOWLEDGMENTS

This material is based on work by C.M.G. supported by the National Science Foundation Postdoctoral Research Fellowship in Biology under Grant No. 1523861. This material is also based on work by K.A.K. supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1650441 and Plant Genome Research Project No. IOS-1238014. This work was supported by the USDA–ARS.

## REFERENCES

Bergquist, R.R. 1981. Transfer from *Tripsacum dactyloides* to corn of a major gene locus conditioning resistance to *Puccinia sorghi*. Phytopathology 71:518. doi:10.1094/Phyto-71-518

Berhan, A.M., S.H. Hulbert, L.G. Butler, and J.L. Bennetzen. 1993. Structure and evolution of the genomes of *Sorghum bicolor* and *Zea mays*. Theor. Appl. Genet. 86:598–604. doi:10.1007/BF00838715

Bomblies, K., and J.F. Doebley. 2005. Molecular evolution of *FLORICAULA/LEAFY* orthologs in the Andropogoneae (Poaceae). Mol. Biol. Evol. 22:1082–1094. doi:10.1093/molbev/msi095

Branson, T.F. 1971. Resistance in the grass tribe Maydeae to larvae of the Western Corn Rootworm. Ann. Entomol. Soc. Am. 64:861–863. doi:10.1093/aesa/64.4.861

Brohammer, A.B., T.J.Y. Kono, N.M. Springer, S.E. McGaugh, and C.N. Hirsch. 2018. The limited role of differential fractionation in genome content variation and function in maize (*Zea mays* L.) inbred lines. Plant J. 93:131–141. doi:10.1111/tpj.13765

Carter, P.R. 1995. Late spring frost and postfrost clipping effect on corn growth and yield. J. Prod. Agric. 8:203–209. doi:10.2134/jpa1995.0203

Chaudhary, B., L. Flagel, R.M. Stupar, J.A. Udall, N. Verma, N.M. Springer, and J.F. Wendel. 2009. Reciprocal silencing, transcriptional bias and functional divergence of homeologs in polyploid cotton (Gossypium). Genetics 182:503–517. doi:10.1534/genetics.109.102608

Chia, J.M., C. Song, P.J. Bradbury, D. Costich, N. De Leon, J. Doebley, R.J. Elshire, B. Gaut, L. Geller, J.C. Glaubitz, et al. 2012. Maize HapMap2 identifies extant variation from a genome in flux. Nat. Genet. 44:803–807. doi:10.1038/ng.2313

Choi, K., X. Zhao, K.A. Kelly, O. Venn, J.D. Higgins, N.E. Yelina, et al. 2013. Arabidopsis meiotic crossover hot spots overlap with H2A.Z nucleosomes at gene promoters. Nat. Genet. 45:1327–1336. doi:10.1038/ng.2766

Clément, Y., G. Sarah, Y. Holtz, F. Homa, S. Pointet, S. Contreras, B. Nabholz, F. Sabot, L. Sauné, M. Ardisson, et al. 2017. Evolutionary forces affecting synonymous variations in plant genomes. PLoS Genet. 13:e1006799. doi:10.1371/journal.pgen.1006799

Conant, G.C., J.A. Birchler, and J.C. Pires. 2014. Dosage, duplication, and diploidization: Clarifying the interplay of multiple models for duplicate gene evolution over time. Curr. Opin. Plant Biol. 19:91–98. doi:10.1016/j.pbi.2014.05.008

deWet, J.M.J., J.R. Harlan, and D.E. Brink. 1982. Systematics of *Tripsacum dactyloides* (Gramineae). Am. J. Bot. 69:1251–1257. doi:10.1002/j.1537-2197.1982.tb13370.x

Dobin, A., C.A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T.R. Gingeras. 2012. STAR: Ultrafast universal RNA-seq aligner. Bioinformatics 29:15–21. doi:10.1093/bioinformatics/bts635

Eddy, S.R. 2009. A new generation of homology search tools based on probabilistic inference. Genome Inform. 23:205–211.

Elmore, R.W., and B. Doupnik. 1995. Corn recovery from early-season frost. J. Prod. Agric. 8:199–203. doi:10.2134/jpa1995.0199

Finn, R.D., P. Coggill, R.Y. Eberhardt, S.R. Eddy, J. Mistry, A.L. Mitchell, et al. 2016. The Pfam protein families database: Towards a more sustainable future. Nucleic Acids Res. 44:D279–D285. doi:10.1093/nar/gkv1344

Freeling, M. 2009. Bias in plant gene content following different sorts of duplication: Tandem, whole-genome, segmental, or by transposition. Annu. Rev. Plant Biol. 60:433–453. doi:10.1146/annurev.arplant.043008.092122

Garsmeur, O., J.C. Schnable, A. Almeida, C. Jourda, A. D'Hont, and M. Freeling. 2014. Two evolutionarily distinct classes of paleopolyploidy. Mol. Biol. Evol. 31:448–454. doi:10.1093/molbev/mst230

Grabherr, M.G., B.J. Haas, M. Yassour, J.Z. Levin, D.A. Thompson, I. Amit, et al. 2011. Full-length transcriptome assembly from RNA-seq data without a reference genome. Nat. Biotechnol. 29:644–652. doi:10.1038/nbt.1883

Gurney A.L., D. Grimanelli, F. Kanampiu, D. Hoisington, J.D. Scholes, and M.C. Press. 2003. Novel sources of resistance to *Striga hermonthica* in *Tripsacum dactyloides*, a wild relative of maize. New Phytol. doi:10.1046/j.1469-8137.2003.00904.x

Haas, B.J., A.L. Delcher, S.M. Mount, J.R. Wortman, R.K. Smith, L.I. Hannick, et al. 2003. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res. 31:5654–5666. doi:10.1093/nar/gkg770

Hellsten, U., K.M. Wright, J. Jenkins, S. Shu, Y. Yuan, S.R. Wessler, J. Schmutz, J.H. Willis, and D.S. Rokhsar. 2013. Fine-scale variation in meiotic recombination in Mimulus inferred from population shotgun sequencing. Proc. Natl. Acad. Sci. USA 110:19478–19482. doi:10.1073/pnas.1319032110

Hirsch, C.N., C.D. Hirsch, A.B. Brohammer, M.J. Bowman, I. Soifer, O. Barad, et al. 2016. Draft assembly of elite inbred line PH207 provides insights into genomic and transcriptome diversity in maize. Plant Cell 28:2700–2714. doi:10.1105/tpc.16.00353

Hughes, T.E., J.A. Langdale, and S. Kelly. 2014. The impact of widespread regulatory neofunctionalization on homeolog gene evolution following whole-genome duplication in maize. Genome Res. 24:1348–1355. doi:10.1101/gr.172684.114

Kumar, S., M. Jones, G. Koutsovoulos, M. Clarke, and M. Blaxter. 2013. Blobology: Exploring raw genome data for contaminants, symbionts and parasites using taxon-annotated GC-coverage plots. Front. Genet. 4:237. doi:10.3389/fgene.2013.00237

Langmead, B., and S.L. Salzberg. 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9:357–359. doi:10.1038/nmeth.1923

Law, M., K.L. Childs, M.S. Campbell, J.C. Stein, A.J. Olson, C. Holt, et al. 2015. Automated update, revision, and quality control of the maize genome annotations using MAKER-P improves the B73 RefGen_v3 gene models and identifies new genes. Plant Physiol. 167:25–39. doi:10.1104/pp.114.245027

Li, Z., G. Hu, X. Liu, Y. Zhou, Y. Li, X. Zhang, et al. 2016. Transcriptome sequencing identified genes and gene ontologies associated with early freezing tolerance in maize. Front. Plant Sci. 7:1477. doi:10.3389/fpls.2016.01477

Lynch, M., and J.S. Conery. 2000. The evolutionary fate and consequences of duplicate genes. Science 290:1151–1155. doi:10.1126/science.290.5494.1151

Lynch, M., and J.S. Conery. 2003. The evolutionary demography of duplicate genes. J. Struct. Funct. Genomics 3:35–44. doi:10.1023/A:1022696612931

Maere, S., S. De Bodt, J. Raes, T. Casneuf, M. Van Montagu, M. Kuiper, and Y. Van de Peer. 2005. Modeling gene and genome duplications in eukaryotes. Proc. Natl. Acad. Sci. USA 102:5454–5459. doi:10.1073/pnas.0501102102

Mathews, S., R.E. Spangler, R.J. Mason-Gamer, and E.A. Kellogg. 2002. Phylogeny of Andropogoneae inferred from phytochrome B, GBSSI, and *ndhF*. Int. J. Plant Sci. 163:441–450. doi:10.1086/339155

Moellenbeck, D.J., B.D. Barry, and L.L. Darrah. 1995. *Tripsacum dactyloides* (Gramineae) seedlings for host plant resistance to the Western Corn Rootworm (Coleoptera: Chrysomelidae). J. Econ. Entomol. 88:1801–1803. doi:10.1093/jee/88.6.1801

Murat, F., J.H. Xu, E. Tannier, M. Abrouk, N. Guilhot, C. Pont, J. Messing, and J. Salse. 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

Muyle, A., L. Serres-Giardi, A. Ressayre, J. Escobar, and S. Glémin. 2011. GC-biased gene conversion and selection affect GC content in the *Oryza* genus (rice). Mol. Biol. Evol. 28:2695–2706. doi:10.1093/molbev/msr104

Nussbaumer, T., M.M. Martis, S.K. Roessner, M. Pfeifer, K.C. Bader, S. Sharma, H. Gundlach, and M. Spannagl. 2013. MIPS PlantsDB: A database framework for comparative plant genome research. Nucleic Acids Res. 41:D1144–D1151. doi:10.1093/nar/gks1153

Pophaly, S.D., and A. Tellier. 2015. Population level purifying selection and gene expression shape subgenome evolution in maize. Mol. Biol. Evol. 32:3226–3235.

Rice, P., I. Longden, and A. Bleasby. 2000. EMBOSS: The European molecular biology open software suite. Trends Genet. 16:276–277. doi:10.1016/S0168-9525(00)02024-2

Rodgers-Melnick, E., P.J. Bradbury, R.J. Elshire, J.C. Glaubitz, C.B. Acharya, S.E. Mitchell, C. Li, Y. Li, and E.S. Buckler. 2015. Recombination in diverse maize is stable, predictable, and associated with genetic load. Proc. Natl. Acad. Sci. USA 112:3823–3828. doi:10.1073/pnas.1413864112

Ross-Ibarra, J., M. Tenaillon, and B.S. Gaut. 2009. Historical divergence and gene flow in the genus *Zea*. Genetics 181:1399–1413. doi:10.1534/genetics.108.097238

Roulin, A., P.L. Auer, M. Libault, J. Schlueter, A. Farmer, G. May, G. Stacey, R.W. Doerge, and S.A. Jackson. 2013. The fate of duplicated genes in a polyploid plant genome. Plant J. 73:143–153. doi:10.1111/tpj.12026

Sankoff, D., C. Zheng, and Q. Zhu. 2010. The collapse of gene complement following whole genome duplication. BMC Genomics 11:313. doi:10.1186/1471-2164-11-313

Schnable, J.C. 2015. Genome evolution in maize: From genomes back to genes. Annu. Rev. Plant Biol. 66:329–343. doi:10.1146/annurev-arplant-043014-115604

Schnable, J.C., N.M. Springer, and M. Freeling. 2011. Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. Proc. Natl. Acad. Sci. USA 108:4069–4074. doi:10.1073/pnas.1101368108

Serres-Giardi, L., K. Belkhir, J. David, and S. Glémin. 2012. Patterns and evolution of nucleotide landscapes in seed plants. Plant Cell 24:1379–1397. doi:10.1105/tpc.111.093674

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

Singh, R., R. Ming, and Q. Yu. 2016. Comparative analysis of GC content variations in plant genomes. Trop. Plant Biol. 9:136–149. doi:10.1007/s12042-016-9165-4

Soreng, R.J., P.M. Peterson, K. Romaschenko, G. Davidse, F.O. Zuloaga, E.J. Judziewicz, T.S. Filgueiras, J.I. Davis, and O. Morrone. 2015. A worldwide phylogenetic classification of the Poaceae (Gramineae). J. Syst. Evol. 53:117–137. doi:10.1111/jse.12150

Swigonová, Z., J. Lai, J. Ma, W. Ramakrishna, V. Llaca, J.L. Bennetzen, and J. Messing. 2004. Close split of sorghum and maize genome progenitors. Genome Res. 14:1916–1923. doi:10.1101/gr.2332504

Tasdighian, S., M. Van Bel, Z. Li, Y. Van de Peer, L. Carretero-Paulet, and S. Maere. 2017. Reciprocally retained genes in the angiosperm lineage show the hallmarks of dosage balance sensitivity. Plant Cell 29:2766–2785. doi:10.1105/tpc.17.00313

Thibaud-Nissen, F., S. Ouyang, and C.R. Buell. 2009. Identification and characterization of pseudogenes in the rice gene complement. BMC Genomics 10:317. doi:10.1186/1471-2164-10-317

Thomas, B.C., B. Pedersen, and M. Freeling. 2006. Following tetraploidy in an Arabidopsis ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. Genome Res. 16:934–946. doi:10.1101/gr.4708406

Tiley, G.P., J.G. Burleigh, and G. Burleigh. 2015. The relationship of recombination rate, genome structure, and patterns of molecular evolution across angiosperms. BMC Evol. Biol. 15:194. doi:10.1186/s12862-015-0473-3

Walley, J.W., R.C. Sartor, Z. Shen, R.J. Schmitz, K.J. Wu, M.A. Urich, et al. 2016. Integration of omic networks in a developmental atlas of maize. Science 353:814–818. doi:10.1126/science.aag1125

Wang, X., J. Wang, D. Jin, H. Guo, T.H. Lee, T. Liu, and A.H. Paterson. 2015. Genome alignment spanning major Poaceae lineages reveals heterogeneous evolutionary rates and alters inferred dates for key evolutionary events. Mol. Plant 8:885–898. doi:10.1016/j.molp.2015.04.004

Wei, F., E. Coe, W. Nelson, A.K. Bharti, F. Engler, E. Butler, et al. 2007. Physical and genetic structure of the maize genome reflects its complex evolutionary history. PLoS Genet. 3:e123. doi:10.1371/journal.pgen.0030123

Whitkus, R., J. Doebley, and M. Lee. 1992. Comparative genome mapping of Sorghum and maize. Genetics 132:1119–1130.

Wong, G.K.S., J. Wang, L. Tao, J. Tan, J. Zhang, D.A. Passey, and J. Yu. 2002. Compositional gradients in *Gramineae* genes. Genome Res. 12:851–856. doi:10.1101/gr.189102

Woodhouse, M.R., F. Cheng, J.C. Pires, D. Lisch, M. Freeling, and X. Wang. 2014. Origin, inheritance, and gene regulatory consequences of genome dominance in polyploids. Proc. Natl. Acad. Sci. USA 111:5283–5288. doi:10.1073/pnas.1402475111

Woodhouse, M.R., J.C. Schnable, B.S. Pedersen, E. Lyons, D. Lisch, S. Subramaniam, and M. Freeling. 2010. Following tetraploidy in maize, a short deletion mechanism removed genes preferentially from one of the two homologs. PLoS Biol. 8:e1000409. doi:10.1371/journal.pbio.1000409

Xiao, J., M.K. Sekhwal, P. Li, R. Ragupathy, S. Cloutier, X. Wang, and F.M. You. 2016. Pseudogenes and their genome-wide prediction in plants. Int. J. Mol. Sci. 17:1991. doi:10.3390/ijms17121991

Yan L., X. Lai, O. Rodriguez, S. Mahboub, R. Rosten, and J. Schnable. 2018. Parallel natural selection in the cold-adapted crop-wild relative Tripsacum dactyloides and artificial selection in temperate adapted maize. Preprint, submitted 22 July 2018. bioRxiv. doi:10.1101/187575

Zhao, M., B. Zhang, D. Lisch, and J. Ma. 2017. Patterns and consequences of subgenome differentiation provide insights into the nature of paleopolyploidy in plants. Plant Cell 29:2974–2994. doi:10.1105/tpc.17.00595

Zhao, Y., L. Tang, Z. Li, J. Jin, J. Luo, and G. Gao. 2015. Identification and analysis of unitary loss of long-established protein-coding genes in Poaceae shows evidences for biased gene loss and putatively functional transcription of relics. BMC Evol. Biol. 15:66. doi:10.1186/s12862-015-0345-x

Zhu, Q., Z. Cai, Q. Tang, and W. Jin. 2016. Repetitive sequence analysis and karyotyping reveal different genome evolution and speciation of diploid and tetraploid *Tripsacum dactyloides*. Crop J. 4:247–255. doi:10.1016/j.cj.2016.04.003

Zou, C., M.D. Lehti-Shiu, F. Thibaud-Nissen, T. Prakash, C.R. Buell, and S.H. Shiu. 2009. Evolutionary and expression signatures of pseudogenes in Arabidopsis and rice. Plant Physiol. 151:3–15. doi:10.1104/pp.109.140632