Parallels between artificial selection in temperate maize and natural selection in the cold-adapted crop-wild relative *Tripsacum*

Lang Yan[1,2,4,5,+], Xianjun Lai[1,2,3,+], Oscar Rodriguez[2], Samira Mahboub[1,6], Rebecca L. Roston[1,6], and James C. Schnable[1,2*]

[1]Center for Plant Science Innovation, University of Nebraska-Lincoln, Lincoln, 68588, USA

[2]Department of Agronomy and Horticulture, University of Nebraska-Lincoln, Lincoln, 68588, USA

[3]Maize Research Institute, Sichuan Agricultural University, Chengdu, 611130, China

[4]College of Life Sciences, Sichuan University, Chengdu, 610065, China

[5]Laboratory of Functional Genome and Application of Potato, Xichang College, Liangshan, 615000, China

[6]Department of Biochemistry, University of Nebraska-Lincoln, Lincoln, Nebraska, USA

[*]To whom correspondence should be addressed. Tel: +1(402)472-3192; E-mail: schnable@unl.edu

[+]these authors contributed equally to this work

Running head: parallel selection in maize and tripsacum

ABSTRACT

The direct progenitor of maize, teosinte, is indigenous to a relatively small range of tropical and sub-tropical latitudes. In contrast, domesticated maize thrives in temperate climates around the world as a result of artificial selection for adaption to these regions. *Tripsacum*, a sister genus to maize and teosinte, is naturally endemic to almost all areas in the western hemisphere where maize is cultivated. A full-length reference transcriptome for tripsacum generated using long-read isoseq data was used to characterize independent adaptation to temperate climates in these taxa, as well as the shared and lineage specific consequences of whole genome duplicate which occurred in the common ancestor of these taxa. Genes related to phospholipid biosynthesis were enriched among those genes experiencing more rapid rates of protein sequence evolution in tripsacum than in other grass lineages. In contrast with previous studies of genes under parallel selection during domestication, we find that there is a statistically significant overlap in the genes which were targets of artificial selection during the adaptation of maize to temperate climates and were targets of natural selection in temperate adapted tripsacum. The overlap between the targets of natural and artificial selection suggests genetic changes in crop-wild relatives associated with adaptation to new environments may be useful guides for identifying genetic targets for breeding efforts aimed at adapting crops to a changing climate.

Keywords: Tripsacum, Maize, full-length transcriptome, natural selection, artificial selection, climate adaptation

Introduction

As both a leading model for plant genetics and one of the three crops that provides 1/2 of all calories consumed by humans around the world, maize (*Zea mays* ssp. *mays*) and its wild relatives have been the subject of widespread genetic and genomic investigations. Biodiversity from wild relatives is a powerful tool which can be utilized in crop breeding efforts (Stepp et al., 2002). The closest relatives of maize are the teosintes, which include the direct wild progenitor of the crop (*Z. mays* ssp. *parviglumis*) as well as a number of other teosinte species within the genus *Zea* (Table 1). Outside the genus *Zea*, the closest relatives of maize are the members of the sister genus *Tripsacum* (Figure 1A-H). Together, these two genera form the subtribe *Tripsacinae* within the tribe *Andropogoneae* (Group et al., 2001). Despite their close genetic relationship, tripsacums are adapted to a much wider range of climates than wild teosintes (Figure 1).

The direct progenitor of maize, *Z. m. parviglumis*, is confined to a relatively narrow native range that spans tropical and subtropical areas of Mexico, Guatemala, Nicaragua and Honduras (Iltis and Doebley, 1980). Other species and subspecies within the genus *Zea* are also largely confined to the same geographic region (Doebley, 1990; Iltis and Doebley, 1980). In contrast, species from the genus *Tripsacum* are widely distributed through temperate regions of both North and South America (Doebley, 1983; Forster and Kole, 2011), largely mirroring the distribution of modern agricultural production of maize in the western hemisphere (Figure 1). The common ancestor of *Zea* and *Tripsacum* is predicted to have been adapted to tropical latitudes (Doebley, 1983; Edwards and Smith, 2010). Therefore, the study of how natural selection adapted tripsacum to temperate climates is a potentially informative parallel to the study of how artificial selection adapted maize to temperate climates. Tripsacum is also a potential source of insight into the genetic changes responsible for traits such disease and insect

resistance, drought and frost tolerance, many of which are targets for maize improvement (Chia et al., 2012; Wang et al., 2013a).

Comparative genetic mapping in maize and tripsacum has demonstrated significant conservation of synteny between the two species (Blakey, 1993) which is also supported by studies of a translocation of tripsacum sequence onto maize chromosome 2 (Maguire, 1962) and studies of tripsacum addition lines of maize (Galinat, 1973) which demonstrated tripsacum carried genomic blocks with dominant loci able to complement multiple genetically linked recessive maize mutants. The whole genome duplication (WGD) present in maize (Schnable et al., 2011; Swigoňová et al., 2004) occurred before the split of the *Zea* and *Tripsacum* lineages (Bomblies and Doebley, 2005; Chia et al., 2012). Fractionation of duplicate genes from the *Zea-Tripsacum* WGD has been shown to be ongoing in the maize lineage, with some retained gene copies present in some maize haplotypes by missing from others (Hirsch et al., 2016; Schnable et al., 2011). Therefore it is likely that homeologous regions in the tripsacum genome may have experienced some degree of reciprocal gene loss events relative to gene loss events present in maize and thus contain ancestral maize genes lost from the maize lineage over the last several million years. The date of the *Zea-Tripsacum* WGD has been variously estimated to be 11.4 million years ago (mya) (Gaut and Doebley, 1997), 4.8-11.9 mya (Swigoňová et al., 2004), and 25.8 mya (Wang et al., 2015). Estimates of the date of divergence between *Zea* and *Tripsacum* or the age of this split relative to their shared WGD have been less common, likely as a result of the comparative dearth of tripsacum genetic and genomic resources (Blakey et al., 2007).

Existing genetic and genomic resources for tripsacum have largely been generated as outgroups for molecular evolution studies in maize (as reviewed (Blakey et al., 2007)). As part of Hapmap2, 8x short read shotgun data generated from tripsacum (Chia et al., 2012), and additional low pass genomic data has been generated for several other tripsacum species (Zhu et al., 2016). Here we applied PacBio long-read sequencing to generate a set of full length transcript sequences from *Tripsacum dactyloides*. Data was generated using RNA isolated from three vegetative tissues to increase the overall fraction of expressed transcripts sampled. The same technology has been employed to generate full length transcript sequences in maize (Wang et al., 2016) enabling comparisons of transcript isoforms based on full length reads. Previous analyses based on short read data found that syntenic genes are more than twice as likely to exhibit conserved alternative splicing patterns as nonsyntenic genes (Mei et al., 2017a) but that in some cases alternative splicing had diverged between maize homeologs with one copy retaining an ancestral splicing pattern shared with sorghum (Mei et al., 2017b). Using data from orthologous genes in maize, tripsacum, sorghum, setaria, and oropetium, a set of genes with uniquely high rates of nonsynonymous substitution in tripsacum were identified. We show that these genes are enriched among genes which were also targets of selection during the adaption of domesticated maize to temperate climates through artificial selection, and that a metabolic pathway identified through this method – phospholipid metabolism – has plausible links to cold and freezing tolerance and shows functional divergence between maize and tripsacum.

Results

Full-length tripsacum transcriptome sequencing

A single *Tripsacum dactyloides* plant grown from seed collected from the wild in eastern Nebraska (USA) was used as the donor for all RNA samples. RNA extracted from three tissues (root, leaf, and stem) was used to construct three size

fractionated libraries (1-2, 2-3, and 3-6 kb) which were sequenced using a PacBio RS II yielding a total of 532,071 Read Of Inserts (ROIs). The SMRT Pipe v2.3.0 classified more than half (267,186, 50.2%) of the ROIs as full-length and non-chimeric (FLNC) transcripts based on the presence of 5'-, 3'-cDNA primers and polyA tails. Each size-fractionated library had expected average length of FLNC transcripts of 1,364 bp, 2,272 bp, and 3,323 bp, with the average length of total FLNC transcripts of 2,354 bp, ranging from 300 to 29,209 bp (Table S1). ICE and Quiver processing of FLNC transcripts produced a total of 64,362 high quality (HQ) consensus transcript sequences with an estimated consensus base call accuracy $\geq$ 99% (Figure S1).

Final consensus tripsacum sequences were mapped to the maize reference genome (RefGen_v3) using GMAP. Consistent with previously reported low overall rate of divergence in gene content and gene sequence between maize and tripsacum (Chia et al., 2012), 98.04% (63,103 out of 64,362 HQ consensus sequences) could be confidently mapped to the maize reference genome. Pbtranscript-TOFU was used to collapse the consensus sequences into 24,616 unique isoforms, in which differences in the 5' end of the first exon were considered redundant, and otherwise identical isoforms were merged (see Methods). This final set of unique tripsacum isoforms mapped to a total of 13,089 maize genes, including 7,633 maize genes represented by a single tripsacum transcript and 5,456 maize genes represented by two or more transcripts. Among maize genes to which two or more tripsacum isoforms were mapped the average was 3.1 isoforms per gene. Eighty-four maize genes were represented by more than 10 or more tripsacum isoforms, and the single maize gene represented by the most isoforms was GRMZM2G306345, which encodes a pyruvate, phosphate dikinase (PPDK) protein involved in the fixation of carbon dioxide as part of the C4 photosynthetic pathway, with 83 identified tripsacum isoforms (Table S2). A set of 249 confident lncRNAs were identified among the final tripsacum consensus sequences (See Methods) (Figure 2G). With an average length of 1.45 kb (ranging from 0.51 to 3.5 kb), the distribution of lncRNA sequences is notably larger than the average length of the maize lncRNAs identified using pacbio isoseq 0.67 kb (ranging from 0.2–6.6 kb) (Figure S2) (Wang et al., 2016). Only 17 of these 249 lncRNAs exhibited high sequence similarity (identity $>$ 80%) with lncRNA sequences identified in maize.

In total, 12,826 out of 13,089 tripsacum transcripts mapped to 14,401 annotated maize gene models (Figure 2A,B). In some cases a single consensus tripsacum sequence spanned two or more maize gene models (Figure S3). Maize genes were sorted based on their expression levels using an existing short read RNA-seq dataset from maize seedling tissue (Zhang et al., 2017). Of the 13,089 most highly expressed maize genes, 8,191 (62.6%) were aligned with at least one tripsacum transcript indicating that the expression level of a gene in maize is a relatively good predictor of how likely that gene was to be captured in the tripsacum isoseq data. Because genes located in the chromosome arms of maize tend to exhibit higher levels of expression than genes in pericentromeric regions, this sample of tripsacum genes is likely depleted in the types of genes which are over represented in pericentromic regions. Figure 2C and D illustrate the comparative densities of highly expressed maize genes aligned to tripsacum isoseq transcripts across the 10 chromosomes of maize. Tripsacum transcript density was correlated with the density of maize highly expressed genes (Spearman correlation coefficient r = 0.855, $p <$ 2.2e-16). Among the 12,826 tripsacum transcripts mapped to maize gene models, 11,910 were unique one-to-one mappings. Data on conserved maize-sorghum orthologous gene pairs was used to increase the confidence of these mappings (Figure 2E), resulting in a final set of 9,112 putative sorghum-maize-tripsacum orthologous gene groups (available on `https://figshare.com/s/6d55867b09e014eb7aed`, Figure 2F). These gene pairs were grouped into three categories "one-to-one", single tripsacum/maize orthologs without duplication (5,641

gene pairs); "one-to-two", a single tripsacum gene mapped to one copy of a homeologous maize gene pair (1,964 gene triplet); "two-to-two", unique tripsacum sequences mapped to each copy of a homeologous maize gene pair (1,507 gene quartets, Figure 2H).

Tripsacinae pan-transcriptome

A total of 1,259 high quality consensus sequences from tripsacum failed to map to the maize reference genome. These sequences were first aligned to NCBI's RefSeq plant database. Two-hundred-and-sixty-three of these sequences sequences aligned to genes from other grass species such as *Sorghum bicolor*, *Setaria italica*, *Oryza sativa*. These genes may either represent genes missed when generating the maize reference genome assembly (Lai et al., 2010), genes present within the maize pan-genome but absent from the specific line used to generate the maize reference genome (Hirsch et al., 2014), or genes present in the common ancestor of maize and tripsacum but lost from the maize lineage sometime after the *Zea/Tripsacum* split. These tripsacum transcripts included a number of genes annotated as encoding disease resistance proteins and receptor-like protein kinases in other grasses (Table 2). These sequences were aligned to the RefSeq databases for bacteria, fungi, and viruses to rule out contamination as a source for the remaining unaligned reads. Only 3 sequences were identified, all of which aligned to fungi. We speculate the remaining 993 unaligned tripsacum sequences most likely represented reads with higher error rates, but may also represent extremely fast evolving genes in tripsacum or the emergence of novel genic sequences.

A total of 223 transcripts aligned to the maize genome in locations where no maize gene model was present. After additional validation and QC (see Methods), 94 of these cases appeared to be unannotated maize genes which were supported by both an aligned tripsacum transcript and short read RNA-seq data in maize, and another 102 cases appeared to be genomic sequences conserved between maize and tripsacum which were transcribed in tripsacum but lacked expression evidence in maize. More than two thirds of the 94 potentially unannotated maize genes could be confidently aligned to annotated coding sequences in the reference genomes of sorghum (63%) or setaria (66%). In contrast, the sequences present in both the tripsacum and maize genomes but expressed only in tripsacum were much less likely to be present in outgroup species (sorghum (25%) and setaria (30%)), suggesting the transcription of these genomic sequences may be a comparatively recent feature, potentially unique to the tripsacum lineage.

Trans AS divergence more common than *cis*

Consistent with observations from other plant systems (Akhunov et al., 2013; Barbazuk et al., 2008; Marquez et al., 2012), the most common single AS variants in tripsacum were – in descending order of frequency – intron retention (IntronR), exon skipping (ExonS), alternative donor (AltD), alternative acceptor (AltA), and mutually exclusive exons (MXEs). The remaining isoforms incorporated two or more types of changes and were classified as Complex AS (CompAS) (Figure S4A). While a comparable maize dataset generated using the same long read technology contained more total splicing events, likely as a result of approx 4.5x greater sequencing depth (Wang et al., 2016), the overall proportions of AS events belonging to different categories in maize and tripsacum were more similar to each other than either was to sorghum (Figure S4B). Shifting from overall frequency to the conservation of specific splicing events – defined as identical AS codes at the same physical positions when tripsacum and maize full length transcripts are aligned to the maize reference genome – a total of 4,324 (35.3%) tripsacum

AS events associated with 1,065 genes were also identified in maize (Figure S4C) and more than two third (656, 61.6%) of the conserved AS genes were observed in orthologous gene groups while 409 genes were *Zea-Tripsacum* linage-specific in *Tripsacinae*. Despite greater sequencing depth and sampling a wider range of tissues, the maize Iso-seq dataset did not identify alternative splicing events corresponding to events identified in tripacaum in 85.7% of cases (2,447 of 2,856 orthologous genes). This result is consistent with the rapid divergence of most AS patterns between even closely related species (Figure 3A).

Divergence in AS between orthologous genes in related species can result from either *cis*-regulatory changes associated with the gene undergoing alternative splicing which should therefore create changes unique to a particular gene, or *trans*-regulatory changes which alter the function or expression of splicing factors and which should therefore result in changes in the pattern of AS for different genes distributed across the genome (Reddy et al., 2012; Wang and Burge, 2008). The shared WGD in the *Tripsacum* and *Zea* lineage provides an opportunity to test the relative importance of *cis* and *trans* factors in the gain and/or loss of alternative splicing patterns. Of the 1,542 gene quartets where a single gene in sorghum is co-orthologous to two maize genes and each maize gene is orthologous to a single tripsacum gene, 212 genes exhibited alternative splicing for both tripsacum genes and 409 genes exhibited alternative splicing for both maize genes. In 52.06% of of gene pairs in tripsacum the pattern of splicing was conserved between homeologous gene pairs as well as in 77.8% of maize homeologous gene pairs. Only 57 gene quartets exhibited alternative splicing for both gene copies in both tripsacum and maize. Conserved splicing was more common between either both maize gene copies or both tripsacum gene copies than between orthologous maize-tripsacum gene pairs, despite the fact that maize-tripsacum gene pairs diverged from each other more recently. When the dataset was constrained to the set of cases where two patterns of AS were observed across the four gene copies, shared splicing patterns between both gene copies in the same species remained more common than shared patterns of splicing between orthologous gene pairs in tripsacum and maize (Figure 3B).

Tripsacum genes experiencing rapid protein sequence evolution

A total of 6,950 groups of orthologous genes present in seven grass species were identified using the tripsacum-maize orthologous relationships described above, plus an existing dataset of syntenic orthologous genes across six grass species with known phylogenetic relationships (Figure 4A) (Schnable et al., 2016). These groups included consisted of 4,162 one-to-one, 1,436 one-to-two and 1,352 two-to-two orthologous gene sets due to the WGD event shared by maize and tripsacum. The overall distribution of Ks values for branches leading to individual species scaled with branch length (Figure 4B). However, although maize and tripsacum are sister taxa in the tree used for this analysis, the average maize gene showed more synonymous substitutions than the average tripsacum gene. The split of the *Tripsacum* and *Zea* genera appears to be significantly older than previously estimated (Hilton and Gaut, 1998), at perhaps 7.7 mya (Figure S5) if a widely used rate constant of 6.5×10^{-9} substitutions per synonymous site per year (Gaut et al., 1996) is correct, or 17.3 mya if a recently reported estimated for the divergence of maize and sorghum at 25.8 mya is correct (Wang et al., 2015).

In 2,228 cases the branch leading to tripsacum had the highest Ka/Ks ratio of all branches examined and in 1,817 the branch leading to maize had the highest Ka/Ks ratio (Figure 4C, Figure S6). This bias towards more gene groups showing the highest (ω) values in maize or tripsacum rather than sorghum, setaria, or oropetium, as well as the presence of more extremely

high outlier values in these two species (Figure 4C), is expected as Ka/Ks ratios is based on a smaller absolute counts of substitutions per gene along shorter branches and should therefore exhibit greater variance. The set of genes experiencing accelerate rates of protein sequence evolution in maize were used as a control set to correct for any bias in gene function introduced by these factors.

Genes with signatures of rapid evolution in tripsacum tended to be associated with stress response and glycerophospholipid metabolic process, whereas fast-evolving genes in maize were generally related to microtubule cytoskeleton organization, nutrient reservoir activity and ATPase activity. Figure 5A illustrates the distribution of Ka/Ks ratios in tripsacum and maize respectively for genes where the branch leading to one of these two species exhibited the highest Ka/Ks ratio among the five species examined. Multiple fast-evolving genes involving in cell response to stimulus and stress had extremely high Ka/Ks ratios (> 1) in tripsacum, consistent with positive selection for increased abiotic stress tolerance in tripsacum relative to maize and other related grasses. The annotated functions of the maize orthologs of tripsacum genes experiencing accelerated protein sequence evolution include cold-induced protein, drought-responsive family protein, salt tolerance family protein, etc. (Table 3).

Domesticated maize and wild tripsacum both grow in large temperate regions of the globe, while their common ancestor was adapted to a more tropical environment. The adaptation of tropical maize landraces to temperate environments required changes to flowering time regulation (Swarts et al., 2017) and adaption to new abiotic and biotic stresses. Previous studies have identified a large set of maize genes which were targets of artificial selection during the process of adaptation to temperate climates (Hufford et al., 2012). We hypothesized that the more ancient process of the expansion of tripsacum into temperate climates may have targeted some of the same genes targeted by artificial selection during the introduction of maize into temperate climates. Tripsacum genes were divided into those where the maize ortholog was identified as likely under selection during the adaption of maize to temperate climates and those where the maize ortholog did not show evidence of being under selection during this process. As Ka/Ks ratios can vary widely across different genes as a result of factors including expression level, gene functional category, and location relative to centromeres (Yang and Gaut, 2011) all tripsacum Ka/Ks values were normalized relative to sorghum, a closely related that is still primarily adapted to tropical latitudes. Genes under selection during the development of temperate maize lines showed significant increases in Ka/Ks values in temperate adapted tripsacum relative to tropical adapted sorghum (log transformed t-test, p-value = 0.027) (Figure 5B) (He et al., 2017). This observation remained significant when using the median Ka/Ks value from orthologs in three different tropically adapted grass species (rice, oropetium and sorghum) (long transformed t-test, p-value = 0.038).

Accelerated evolution of phospholipid metabolism in tripsacum

In the process of detecting tripsacum genes that might experience accelerate evolution for temperate climate adaptation, we noticed multiple genes annotated as participating in the phospholipid metabolism pathway where the highest Ka/Ks ratio for that gene observed in the branch leading to tripsacum. While several genes in this pathway also showed signs of accelerated evolution in maize, the bias towards high rates of protein sequence change in tripsacum was dramatic (Figure 5A). Using log-transformed Ka/Ks values genes in the phospholipid biosynthesis pathway exhibited a significantly higher range of p-values

than the background set of other genes (p-value = 3.95e-04). In contrast, maize genes in the same exhibited significantly lower p-values than background maize genes (p-value = 3.08e-03). Comparing the ratio of Ka/Ks values for between same genes in both maize and tripsacum, genes in the phospholipid biosynthesis pathway showed significantly higher ratios of Ka/Ks values than background genes (p-value = 4.253e-05) (Figure S7A).

Phospholipids are a class of lipids that are a major component of cell membranes and include lipids with head groups such as phosphatidate (PA), phosphatidylethanolamine (PE), phosphatidylcholine (PC), phosphatidylglycerol(PG) and phosphatidylserine (PS) which share overlapping biosynthesis pathways and are often inter-convertible (Figure 6). The set of genes involved in phospholipid metabolism which experienced accelerated evolution in tripsacum were particularly concentrated in the pathway leading to PC and as well as a gene encoding the enzyme (1-acylglycerol-3-phosphate O-acyltransferase, EC 2.3.1.51) which synthesizes phosphatidate (PA). These have the potential to be particularly relevant changes because PA is the precursor to all other lipids (Li-Beisson et al., 2013), while PC is the major phospholipid component of the non-plastid membranes and also tends to control membrane desaturation through acyl-editing (Bates et al., 2007). Genes encoding enzymes leading to the primary lipid storage molecule triacylglycerol (TAG) also showed signs of elevated protein substitution rates in tripsacum. Changes in membrane lipid composition are likely required for changes in cold and freezing tolerance, as numerous lipid changes have been documented to prevent or limit damage to cell membranes at cold or freezing temperatures in a number of species (Moellering et al., 2010; Scotti-Campos et al., 2014) and damage to membrane structure and function is one of the primary mechanisms of cell death during freezing stress (Xin and Browse, 2000). Testing confirmed that Tripsacum seedlings grown from seed collected as part of the same expedition were able to tolerate prolonged 4 °C cold stress while the same temperature stress treatment produced significant levels of cell death in maize seedlings from the reference line B73 (Figure S7B-F).

Discussion

The potential for data from *Tripsacum* to aid in both basic biological research and applied plant breeding in maize has long been discussed (Chia et al., 2012; Wang et al., 2013a). However, prior to this project, a total of only 611 published nucleotide sequences existed for the entire *Tripsacum* genus, including 565 for *T. dactyloides*, 12 for *T. andersonii*, and 34 for all other named taxa within the genus. Here we have generated a set of 24,616 full length tripsacum cDNA sequences covering 22.4%, 31.5% and 60.2% of the annotated, expressed, and syntenically conserved gene space of maize respectively. This larger scale transcriptome resource enables a number of comparative analyses of *Zea* and *Tripsacum* not previously feasible.

Significant evolutionary rate heterogeneity exists among extant grass species. Previously, variation in the rate of divergence between homeologous gene pairs generated during the rho polyploidy (Paterson et al., 2004) in different grass species was employed to detect variation in synonymous substitution rates (Wang et al., 2015). However, this approach, which relies on pairwise comparisons between species, provides aggregate estimates for each lineage across the 70-96 million years since the rho WGD (Paterson et al., 2004; Wang et al., 2015). Utilizing known phylogenetic relationships across relatively large numbers of grass species with sequenced genomes or significant genome resources and fitting rates of synonymous and nonsynonymous substitutions for each branch separately (Yang, 1997) demonstrated that even comparing sister genera (*Zea* and *Tripsacum*),

maize exhibits significantly more rapid accumulation of synonymous substitutions. One potential explanation is growth habit. Many *Zea* species are annuals (Doebley, 1990) while more than 13 tripsacum species are perennials (Doebley, 1983) (Table 1) and several analyses have suggested that synonymous substitutions accumulate more rapidly in annual species (Gaut et al., 2011; Smith and Donoghue, 2008). Another potential explanation is that the difference in the rate of molecular evolution between maize and tripsacum may reflect the difference in native range between these genera, as species native tropical regions have been shown to accumulate nucleotide substitutions up to twice as rapidly as temperate species (Wright et al., 2006). The availability of a large number of full length tripsacum CDS sequences also provided an opportunity to revisit the dating of evolutionary events previously estimated using only data from small numbers of genes (Hilton and Gaut, 1998; Swigoňová et al., 2004). Using data from seven species and thousands of individual retained syntenic genes, we estimate that the divergence of the *Tripsacum* and *Zea* genera occurred approximately 7.7 mya, significantly earlier than the previously reported estimate of 4.5-4.8 mya for this event using sequence data from the globulin-1 gene and the same evolutionary rate constant employed in this study (Hilton and Gaut, 1998).

Genes are generally considered to show evidence of positive selection if the frequency of nonsynonymous substitutions is significantly higher than that of synonymous substitutions. However, if positive selection is assumed to be episodic rather than constant, elevated Ka/Ks ratios which are less than one can reflect a mixture of positive selection and purifying selection, relaxation of purifying selection, or statistical noise. Episodic positive selection would also be harder to detect on longer branches where the proportion of evolutionary time a gene spends under positive selection decreases relative to the time spent under purifying selection. Background ratios of Ka/Ks can also vary significantly between different genes reflecting differences in chromosome environment, expression level, function, and the presence or absence of different types of duplicate gene copies (Yang and Gaut, 2011). As shown in Figure 4C, the frequency of extreme Ka/Ks ratios decreases as the overall branch length increases. The inclusion of tripsacum breaks up the long branch between maize and sorghum, permitting the identification of genes experiencing either an interval of positive selection alongside ongoing purifying selection or a relaxation of purifying selection. It must be emphasized that the analysis presented here cannot distinguish between true positives (genes showing elevated rates of protein sequence evolution as a result of positive selection or relaxed selection) and false positives (statistical noise) on a single gene level. Instead the focus must be on the differences observed between the functional classes or pathways of genes which exhibited higher rates of protein sequence evolution in maize and tripsacum. While Ka/Ks ratios are not normally distributed, log transformation can create a normal distribution, permitting the use of a t-test to identify populations of genes exhibiting significantly different Ka/Ks between orthologs in different species (He et al., 2017). Here we found that genes involved in phospholipid metabolism and stress response both tended to be experiencing higher rates of protein sequence evolution in tripsacum than in maize or in the other grass species tested.

Recent studies have demonstrated that parallel selection often acts on largely unlinked sets of genes at a molecular level (Gaut, 2015; Takuno et al., 2015; Lai et al., 2017). This suggests that there are many different molecular mechanisms which can be employed to achieve the same phenotypic changes, and as a result the same genes will rarely be targeted in independent instances of selection for the same traits. In contrast to these previous studies, the evidence presented above suggests that as the lineage leading to *Tripsacum dactyloides* expanded into temperate environments millions of years ago natural selection

targeted some of the same genetic loci which would later be targets of artificial selection as the cultivation of maize spread from the center of domestication in Mexico into more temperate regions of North America. This overlap between targets of natural and artificial selection for adaptation to the same environment in sister genera also indicates that genetic changes in crop-wild relatives associated with adaptation to new environments may be useful guides for identifying genetic targets for breeding efforts aimed at adapting crops to a changing climate.

One specific difference between the native environment of wild *Zea* and *Tripsacum* species is that many *Tripsacum* species grow in areas where they are exposed to cold and freezing temperatures. Unlike maize, *Tripsacum dactyloides* can survive prolonged cold and freezing temperatures and successfully overwinter. The longer growing season enabled by cold tolerance and perenniality has been identified as the major driver responsible for the increased photosynthetic productivity per acre of temperate grass species such as miscanthus relative to maize (Dohleman and Long, 2009). The identification of accelerated protein sequence evolution among genes involved in phospholipid metabolism provides a plausible candidate mechanism for the increased cold and freezing tolerance of tripsacum relative to maize. While widely grown in temperate regions over the summer, maize remains sensitive to cold. Maize varieties with the ability to be planted significantly earlier, or in the extreme case to overwinter, have the potential to intercept a greater proportion of total annual solar radiation increasing maximum potential yields. This study illustrates how studying the genetic mechanisms responsible for crop-wild relative adaptation to particular climates may guide breeding and genome engineering efforts to adapt crops to a changing climate.

Methods

Plant materials and RNA preparation

Tripsacum seeds were collected from wild growing plants located in Eastern Nebraska (USA, GPS coordinates: 41.057836, -96.639844). Seeds with brown cupules were selected. For each seed the cupule was removed, followed by a cold treatment at 4°Cfor at least two days, resulting in germination rates between 30% and 50%. A single plant (ID #LY-1) was selected for transcriptome sequencing. Young leaves were sampled one month after germination. Stem and root were sampled three months after germination. Harvested tissue samples were rinsed with cold distilled water and them immediately frozen in liquid N_2. Total RNA was extracted from each tissue separately by manually grinding each sample in liquid N_2, adding TriPure isolation reagent (Roche Life Science, catalog number #11667157001), followed by separating phase using chloroform, precipitating RNA using isoamyl alcohol and washing the RNA pellet using 75% ethanol. The air-dried RNA samples were dissolved in DEPC-treated water RNA quantity and quality were assessed uisng a a NanoDrop 1000 spectrophotometer and electrophoresis on a 1% agarose gel respectively (Figure S1).

Single-molecule sequencing and isoform detection

Equal quantities of total RNA from each sample were pooled prior to library construction and the combined sample was shipped to the Duke Center for Genomic and Computational Biology (GCB), Duke University, USA for sequencing. Three size-fractionated libraries (1-2 kb, 2-3 kb, and 3-6 kb) were constructed and sequenced separately on the PacBio RS II. Each libraries was sequenced using 2 SMRT cells. Raw reads data was analyzed through running the Iso-Seq pipeline included in the SMRT-

Analysis software package (SMRT Pipe v2.3.0, https://github.com/PacificBiosciences/cDNA_primer).

First, reads of insert consensus sequence (ROIs, previously known as circular consensus sequences, CSCs) were identified using the Reads Of Insert protocol, one of the submodules of the Iso-Seq pipeline. The minimum required number of full passes was set to 0 and predicted consensus accuracy was set to 75%. Raw sequences that passed this step were considered to be the ROIs.

Second, the classify submodule was used to automatically determine which ROIs were full-length. ROIs were considered to be full length if both the 5'- and 3'-cDNA primers as well as a poly(A) tail signal preceding the 3'-primer were detected. The classify submodule also separated chimeric reads from non-chimeric reads through detecting the presence of SMRTbell adapters in the middle of sequences, producing a set of full-length non-chimeric (FLNC) reads.

Third, the isoform-level clustering algorithm ICE (Iterative Clustering for Error Correction), which uses only full length reads, followed by Quiver which polished FLNC consensus sequences from ICE using non-FL reads were used to improve the accuracy of FLNC consensus sequences. These steps resulted in a set of high quality polished consensus with $>$ 99% post-correction accuracy.

Fourth, HQ polished consensus reads were mapped to the maize reference genome (B73_RefGen_v3) using GMAP and redundant mapped transcripts were merged using using the collapse_isoforms_by_sam.py script from the pbtranscript-ToFU package (https://github.com/PacificBiosciences/cDNA_primer/wiki/tofu-Tutorial-(optional).-Removing-redundant-transcripts/) with the parameter settings of min-coverage = 85% and min-identity =82%. Isoforms differing only at the 5'-sites within the first exon were considered to be redundant and collapsed.

Identification of LncRNAs

The protein coding-potential of individual tripsacum transcripts was assessed using CPAT (Coding Potential Assessment Tool) (Wang et al., 2013b), which employs a logistic regression model built with four sequence features as predictor variables: open reading frame (ORF) size, ORF coverage, Fickett testcode statistic, and hexamer usage bias. CPAT was trained using 4,900 high-confidence lncRNAs transcripts identified as part of Gramene 52 release (http://www.gramene.org/release-notes-52/) (Tello-Ruiz et al., 2016) and an equal number of known protein-coding transcripts randomly subsampled from the RefGen_v3 annotation 6a to measure the prediction performance of this logic model in maize. The accuracy of the trained CPAT model was assessed by quantifying six parameters using 10-fold cross validation with the maize training dataset: sensitivity (TPR), specificity (1-FPR), accuracy (ACC) and precision (PPV) under the receiver operating characteristic (ROC) curve and precision–recall (PR) curve. Based on there parameters, a probability threshold of 0.425 was identified as providing the best trade off between specificity and sensitivity for identifying lncRNA sequences.

4,095 candidate non-coding RNAs with a length greater than $>$ 200 bp were predicted in CPAT. ORF prediction for non-coding candidate of tripsacum transcripts was performed using TransDecoder (Haas et al., 2013) and 2,509 transcripts encoding ORFs longer than 100 amino acids were removed from the set of putative lncRNAs. The remaining lncRNAs were aligned to the NCBI-nr database using BLASTX and sequences showing similarity to existing protein sequencing in the database (e-value $\leqslant$ 1e-10) were removed from the set of putative lncRNAs.

Identification of orthologs between tripsacum and maize

Maize-tripsacum orthologs were defined based on the following criteria. Firstly, the tripsacum sequence must have been uniquely mapped to the target maize gene using GMAP. Secondly, the tripsacum gene must have been uniquely mapped to a single sorghum gene using BLASTN. Thirdly, the maize and sorghum genes must be syntenic orthologs of each other using a previously published dataset (Schnable et al., 2016; Zhang et al., 2017). As a result of the maize WGD shared by maize and tripsacum, in some cases multiple tripsacum sequences mapped to unique maize genes on different maize subgenomes, but to the single shared co-ortholog of these two maize genes in sorghum.

BLAST analyses for Isoseq data

Isoforms which failed to map to the maize genome aligned to different NCBI refSeq databases using BLASTN with the following parameters: min-coverage = 50%, min-identity = 70%, max_target_seqs = 5 and e-value $\leqslant$ 1e-10. Transcripts were considered to originate from ancestral grass genes not present in the B73 reference genome if none of the top 5 blast hits were to maize sequences, but did include sequences isolated from other grass species. The transcripts were counted as specifically expressed in tripsacum if the top 5 blast hits were without maize refseq but that in other grasses.

Tripsacum sequences which aligned to regions of the maize genome not annotated as genes were also aligned to the full set of annotated maize gene CDS sequences using BLASTN with parameters: min-coverage=80%, min-identity=85%, max_target_seqs=1 and e-value $\leqslant$ 1e-10. Transcripts identified were considered to be aligned to maize gene models and the remaining ones were then manually proofing using IGV and a set of Illumina short read RNA-seq data from a wide range of maize tissues (Davidson et al., 2011).

Detection of alternative splicing (AS) events

AS analysis was conducted using Astalavista-4.0 (Alternative Splicing Trascriptional Landscape Visualization Too) (Foissac and Sammeth, 2007), transforming identified place isoforms into a set of defined AS codes based on their location within the gene and the type of splicing observed. Among the multiple maize genes with multiple aligned tripsacum isoforms, in 3,566 cases one or more AS events (12,260 in total) was identified between the multiple aligned isoforms, while in the 1,890 remaining cases differences between isoforms were the result of either variation in transcription start sites (VSTs), or additional 3' exons (DSPs). Each AS site is assigned a number according to its relative position in the event and a symbol depending on its type. In addition to the main five single AS types such as Intron retention (IR), Exon skipping (ES), Alternative donor (AD), alternative acceptor (AA) and Mutually exclusive exons (MXEs), other AS events in which multiple types of AS are present between two isoforms were counted as complicated AS. The maize data used to compare AS events between maize and tripsacum was extracted from the published AS dataset generated by Wang et al (Wang et al., 2016).

Substitution rate estimation and selection analyses

Codon based alignments were generated using ParaAT2.0 (Zhang et al., 2012) for sets of orthologous genes identified using a dataset of syntenic orthologous genes identified across six grass species with sequenced genomes (maize V3 (Schnable et al., 2009), sorghum v3.1 (McCormick et al., 2017), setaria v2.2 (Bennetzen et al., 2012), oropetium v1.0 (VanBuren

et al., 2015), rice v7 (Ouyang et al., 2006) and brachypodium v3.1 (Vogel et al., 2010)) (Schnable et al., 2016), plus the maize/tripsacum orthologous relationships defined above. Synonymous nucleotide substitution rates (Ks) were calculated by using the CodeML maximum-likelihood method (runmode = -2, CodonFreq = 2) implemented within PAML (Yang, 1997) and the known phylogenetic relationships of the seven species. The divergence time (T) between maize and tripsacum was estimated following the formula T = Ks/2μ(Gaut et al., 1996).

Branch-specific Ka/Ks ratios with self-built phylogenetic trees were also calculated using Codeml program (runmode = 0). Two models were used, one required a constant value for ω across all branches (model 0), and the other allowed heterogenous rates on each branch of the phylogeny (model 1). A likelihood ratio test was used to compare likelihood values under model 0 and model 1 to test whether significant variation in Ka/Ks ratios between different branches was present (Yang, 1998). For each set of orthologous genes the log likelihood values under two models (lnL1 for the alternative and lnL0 for the null model) were obtained. From these values, the LRT was computed using 2 $\times$ (lnL1-lnL0). A χ^2 curve with degree of freedom = 1 was used to calculate a *p*-value for this LRT. Multiple testing was performed to correct these *p*-values by applying the false discovery rate method (FDR) adjusted in R (Storey and Tibshirani, 2003). A gene was considered to be experiencing accelerated rates of protein evolution in a given lineage if the highest Ka/Ks ratio was detected in the branch leading to that species and the FDR-adjusted *p*-value for the comparison of the constant ω and heterogeneous models was < 0.05.

Use of t-tests in comparisons of Ka/Ks ratios between different populations of genes

Like gene expression fold-change data, Ka/Ks ratios and ratios of Ka/Ks ratios exhibit non-normal distributions. Applying a log transformation to either Ka/Ks ratios or ratios of ratios produces roughly normal distributions, as observed for gene expression fold change data (He et al., 2017). This method was employed to conduct two comparisons using the independent-samples t-test package within R. In the first test, the distribution of ratio of ratio values between tripsacum and sorghum Ka/Ks values was compared between genes identified as likely to be under selection in a comparison of tropical and temperate maize lines (Hufford et al., 2012; Lai et al., 2017) and genes not identified as likely to be under selection in the same comparison. The second test compared Ka/Ks ratios for genes annotated as involved in lipid metabolism to the population of genes not involves in phospholipid metabolism in tripsacum and maize, respectively, as well as Ka/Ks ratios between tripsacum and maize.

Data available

Raw sequencing data are available through the NCBI (`http://www.ncbi.nlm.nih.gov/bioproject`) under accession number SRP114506. Clean data with tripsacum non-redundant isoforms are available at Zenodo with the identifier `http://dx.doi.org/10.5281/zenodo.841005`. All other data are available upon request.

Acknowledgements

We thank Dr. Yang Zhang for the RNA extraction protocol and Dr. Christy Gault (Cornell University) for sharing her protocol for germinating tripsacum seeds. This work was supported by the National Science Foundation under Grant No.OIA-1557417 to JCS, USDA NIFA award 2016-67013-24613 to RLR and JCS, a Science Foundation of Xichang College awarded to LY and a China Scholarship Council fellowship awarded to XL.

Author contributions

JCS, RLR, LY and XL conceived the project and designed the studies; OR identified and collected the plant material used in this study; LY and SM performed the experiments; LY and XL analyzed the data; LY and JCS wrote the paper. All authors reviewed and approved the final manuscript.

References

Akhunov, E.D., Sehgal, S., Liang, H., Wang, S., Akhunova, A.R., Kaur, G., Li, W., Forrest, K.L., See, D., Šimková, H., et al. (2013). Comparative analysis of syntenic genes in grass genomes reveals accelerated rates of gene structure and coding sequence evolution in polyploid wheat. Plant physiology **161**:252–265.

Barbazuk, W.B., Fu, Y., and McGinnis, K.M. (2008). Genome-wide analyses of alternative splicing in plants: opportunities and challenges. Genome research **18**:1381–1392.

Bates, P.D., Ohlrogge, J.B., and Pollard, M. (2007). Incorporation of newly synthesized fatty acids into cytosolic glycerolipids in pea leaves occurs via acyl editing. Journal of Biological Chemistry **282**:31206–31216.

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**:555–561.

Blakey, C., Costich, D., Sokolov, V., and Islam-Faridi, M.N. (2007). Tripsacum genetics: from observations along a river to molecular genomics. Maydica **52**:81.

Blakey, C.A. (1993). A Molecular Map in Tripsacum Dactyloides, Eastern Gamagrass. University of Missouri-Columbia.

Bomblies, K., and Doebley, J.F. (2005). Molecular evolution of floricaula/leafy orthologs in the andropogoneae (poaceae). Molecular Biology and Evolution **22**:1082–1094.

Chia, J.M., Song, C., Bradbury, P.J., Costich, D., de Leon, N., Doebley, J., Elshire, R.J., Gaut, B., Geller, L., Glaubitz, J.C., et al. (2012). Maize hapmap2 identifies extant variation from a genome in flux. Nature genetics **44**:803–807.

Davidson, R.M., Hansey, C.N., Gowda, M., Childs, K.L., Lin, H., Vaillancourt, B., Sekhon, R.S., de Leon, N., Kaeppler, S.M., Jiang, N., et al. (2011). Utility of rna sequencing for analysis of maize reproductive transcriptomes. The Plant Genome **4**:191–203.

De Wet, J., Timothy, D., Hilu, K., and Fletcher, G. (1981). Systematics of south american tripsacum (gramineae). American Journal of Botany 269–276.

Doebley, J. (1983). The taxonomy and evolution of tripsacum and teosinte, the closest relatives of maize. Maize Virus Disease Colloquium and Workshop. The Ohio State University, Ohio Agricultural Research and Development Center, Wooster, Ohio. 15–28.

Doebley, J. (1990). Molecular evidence for gene flow among *Zea* species. BioScience **40**:443–448.

Doebley, J.F., and Iltis, H.H. (1980). Taxonomy of *Zea* (gramineae). i. a subgeneric classification with key to taxa. American Journal of Botany 982–993.

Dohleman, F.G., and Long, S.P. (2009). More productive than maize in the midwest: how does miscanthus do it? Plant

physiology **150**:2104–2115.

Edwards, E.J., and Smith, S.A. (2010). Phylogenetic analyses reveal the shady history of c4 grasses. Proceedings of the National Academy of Sciences **107**:2532–2537.

Eubanks, M.W. (2001). The mysterious origin of maize. Economic Botany 492–514.

Foissac, S., and Sammeth, M. (2007). Astalavista: dynamic and flexible analysis of alternative splicing events in custom gene datasets. Nucleic acids research **35**:W297–W299.

Forster, B., and Kole, C. (2011). Wild crop relatives: Genomic and breeding resources. cereals. Experimental Agriculture **47**:736.

Galinat, W.C. (1973). Intergenomic mapping of maize, teosinte and tripsacum. Evolution **27**:644–655.

Gaut, B., Yang, L., Takuno, S., and Eguiarte, L.E. (2011). The patterns and causes of variation in plant nucleotide substitution rates. Annual Review of Ecology, Evolution, and Systematics **42**:245–266.

Gaut, B.S. (2015). Evolution is an experiment: Assessing parallelism in crop domestication and experimental evolution: (nei lecture, smbe 2014, puerto rico). Molecular biology and evolution **32**:1661–1671.

Gaut, B.S., and Doebley, J.F. (1997). Dna sequence evidence for the segmental allotetraploid origin of maize. Proceedings of the National Academy of Sciences **94**:6809–6814.

Gaut, B.S., Morton, B.R., McCaig, B.C., and Clegg, M.T. (1996). Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene adh parallel rate differences at the plastid gene rbcl. Proceedings of the National Academy of Sciences **93**:10274–10279.

Group, G.P.W., Barker, N.P., Clark, L.G., Davis, J.I., Duvall, M.R., Guala, G.F., Hsiao, C., Kellogg, E.A., Linder, H.P., Mason-Gamer, R.J., et al. (2001). Phylogeny and subfamilial classification of the grasses (poaceae). Annals of the Missouri Botanical Garden 373–457.

Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D., Li, B., Lieber, M., et al. (2013). De novo transcript sequence reconstruction from rna-seq using the trinity platform for reference generation and analysis. Nature protocols **8**:1494–1512.

He, M., Liu, P., and Lawrence-Dill, C.J. (2017). A method to assess significant differences in rna expression among specific gene groups. bioRxiv doi:10.1101/136143.

Hilton, H., and Gaut, B.S. (1998). Speciation and domestication in maize and its wild relatives: evidence from the globulin-1 gene. Genetics **150**:863–872.

Hirsch, C., Hirsch, C.D., Brohammer, A.B., Bowman, M.J., Soifer, I., Barad, O., Shem-Tov, D., Baruch, K., Lu, F., Hernandez, A.G., et al. (2016). Draft assembly of elite inbred line ph207 provides insights into genomic and transcriptome diversity in maize. The Plant Cell Online tpc–00353.

Hirsch, C.N., Foerster, J.M., Johnson, J.M., Sekhon, R.S., Muttoni, G., Vaillancourt, B., Peñagaricano, F., Lindquist, E., Pedraza, M.A., Barry, K., et al. (2014). Insights into the maize pan-genome and pan-transcriptome. The Plant Cell **26**:121–135.

Hufford, M.B., Xu, X., Van Heerwaarden, J., Pyhäjärvi, T., Chia, J.M., Cartwright, R.A., Elshire, R.J., Glaubitz, J.C.,

Guill, K.E., Kaeppler, S.M., et al. (2012). Comparative population genomics of maize domestication and improvement. Nature genetics **44**:808–811.

Iltis, H.H., and Benz, B.F. (2000). Zea nicaraguensis (poaceae), a new teosinte from pacific coastal nicaragua. Novon 382–390.

Iltis, H.H., and Doebley, J.F. (1980). Taxonomy of *Zea* (gramineae). ii. subspecific categories in the *Zea mays* complex and a generic synopsis. American Journal of Botany 994–1004.

Iltis, H.H., Doebley, J.F., Guzmán, R., and Pazy, B. (1979). Zea diploperennis (gramineae): a new teosinte from mexico. Science **203**:186–188.

Lai, J., Li, R., Xu, X., Jin, W., Xu, M., Zhao, H., Xiang, Z., Song, W., Ying, K., Zhang, M., et al. (2010). Genome-wide patterns of genetic variation among elite maize inbred lines. Nature genetics **42**:1027–1030.

Lai, X., Yan, L., Lu, Y., and Schnable, J. (2017). Largely unlinked gene sets targeted by selection for domestication syndrome phenotypes in maize and sorghum. bioRxiv 184424.

Li-Beisson, Y., Shorrosh, B., Beisson, F., Andersson, M.X., Arondel, V., Bates, P.D., Baud, S., Bird, D., DeBono, A., Durrett, T.P., et al. (2013). Acyl-lipid metabolism. The Arabidopsis Book **11**:e0161.

Maguire, M.P. (1962). Common loci in corn and tripsacum. Journal of Heredity **53**:87–88.

Marquez, Y., Brown, J.W., Simpson, C., Barta, A., and Kalyna, M. (2012). Transcriptome survey reveals increased complexity of the alternative splicing landscape in arabidopsis. Genome research **22**:1184–1195.

McCormick, R.F., Truong, S.K., Sreedasyam, A., Jenkins, J., Shu, S., Sims, D., Kennedy, M., Amirebrahimi, M., Weers, B., McKinley, B., et al. (2017). The sorghum bicolor reference genome: improved assembly and annotations, a transcriptome atlas, and signatures of genome organization. bioRxiv 110593.

Mei, W., Boatwright, J.L., Feng, G., Schnable, J.C., and Barbazuk, W.B. (2017a). Evolutionarily conserved alternative splicing across monocots. Genetics doi:10.1534/genetics.117.300189.

Mei, W., Liu, S., Schnable, J.C., Yeh, C.T., Springer, N.M., Schnable, P.S., and Barbazuk, W.B. (2017b). A comprehensive analysis of alternative splicing in paleopolyploid maize. Frontiers in Plant Science **8**:694.

Moellering, E.R., Muthan, B., and Benning, C. (2010). Freezing tolerance in plants requires lipid remodeling at the outer chloroplast membrane. Science **330**:226–228.

Ouyang, S., Zhu, W., Hamilton, J., Lin, H., Campbell, M., Childs, K., Thibaud-Nissen, F., Malek, R.L., Lee, Y., Zheng, L., et al. (2006). The tigr rice genome annotation resource: improvements and new features. Nucleic acids research **35**:D883–D887.

Paterson, A., Bowers, J., and Chapman, B. (2004). Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. Proceedings of the National Academy of Sciences of the United States of America **101**:9903–9908.

Reddy, A.S., Rogers, M.F., Richardson, D.N., Hamilton, M., and Ben-Hur, A. (2012). Deciphering the plant splicing code: experimental and computational approaches for predicting alternative splicing and splicing regulatory elements. Frontiers in plant science **3**:18.

Schnable, J., Zang, Y., and W.C. Ngu, D. (2016). Pan-grass syntenic gene set (sorghum referenced). Figshare

https://dx.doi.org/10.6084/m9.figshare.3113488.v1.

Schnable, J.C., Springer, N.M., and Freeling, M. (2011). Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. Proceedings of the National Academy of Sciences **108**:4069–4074.

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.

Scotti-Campos, P., Pais, I.P., Partelli, F.L., Batista-Santos, P., and Ramalho, J.C. (2014). Phospholipids profile in chloroplasts of coffea spp. genotypes differing in cold acclimation ability. Journal of plant physiology **171**:243–249.

Smith, S.A., and Donoghue, M.J. (2008). Rates of molecular evolution are linked to life history in flowering plants. science **322**:86–89.

Stepp, J.R., Wyndham, F.S., and Zarger, R.K. (2002). Ethnobiology and biocultural diversity: proceedings of the Seventh International Congress of Ethnobiology. University of Georgia Press.

Storey, J.D., and Tibshirani, R. (2003). Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences **100**:9440–9445.

Swarts, K., Gutaker, R.M., Benz, B., Blake, M., Bukowski, R., Holland, J., Kruse-Peeples, M., Lepak, N., Prim, L., Romay, M.C., et al. (2017). Genomic estimation of complex traits reveals ancient maize adaptation to temperate north america. Science **357**:512–515.

Swigoňová, Z., Lai, J., Ma, J., Ramakrishna, W., Llaca, V., Bennetzen, J.L., and Messing, J. (2004). Close split of sorghum and maize genome progenitors. Genome research **14**:1916–1923.

Takuno, S., Ralph, P., Swarts, K., Elshire, R.J., Glaubitz, J.C., Buckler, E.S., Hufford, M.B., and Ross-Ibarra, J. (2015). Independent molecular basis of convergent highland adaptation in maize. Genetics **200**:1297–1312.

Tello-Ruiz, M.K., Stein, J., Wei, S., Youens-Clark, K., Jaiswal, P., and Ware, D. (2016). Gramene: a resource for comparative analysis of plants genomes and pathways. Plant Bioinformatics: Methods and Protocols 141–163.

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.

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

Wang, B., Tseng, E., Regulski, M., Clark, T.A., Hon, T., Jiao, Y., Lu, Z., Olson, A., Stein, J.C., and Ware, D. (2016). Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. Nature communications **7**.

Wang, C., Shi, X., Liu, L., Li, H., Ammiraju, J.S., Kudrna, D.A., Xiong, W., Wang, H., Dai, Z., Zheng, Y., et al. (2013a). Genomic resources for gene discovery, functional genome annotation, and evolutionary studies of maize and its close relatives. Genetics **195**:723–737.

Wang, L., Park, H.J., Dasari, S., Wang, S., Kocher, J.P., and Li, W. (2013b). Cpat: Coding-potential assessment tool using an alignment-free logistic regression model. Nucleic acids research **41**:e74–e74.

Wang, X., Wang, J., Jin, D., Guo, H., Lee, T.H., Liu, T., and Paterson, A.H. (2015). Genome alignment spanning major

poaceae lineages reveals heterogeneous evolutionary rates and alters inferred dates for key evolutionary events. Molecular plant **8**:885–898.

Wang, Z., and Burge, C.B. (2008). Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. Rna **14**:802–813.

Wright, S., Keeling, J., and Gillman, L. (2006). The road from santa rosalia: a faster tempo of evolution in tropical climates. Proceedings of the National Academy of Sciences **103**:7718–7722.

Xin, Z., and Browse, J. (2000). Cold comfort farm: the acclimation of plants to freezing temperatures. Plant, Cell & Environment **23**:893–902.

Yang, L., and Gaut, B.S. (2011). Factors that contribute to variation in evolutionary rate among arabidopsis genes. Molecular Biology and Evolution **28**:2359–2369.

Yang, Z. (1997). Paml: a program package for phylogenetic analysis by maximum likelihood. Bioinformatics **13**:555–556.

Yang, Z. (1998). Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Molecular biology and evolution **15**:568–573.

Zhang, Y., Ngu, D.W., Carvalho, D., Liang, Z., Qiu, Y., Roston, R.L., and Schnable, J.C. (2017). Differentially regulated ortholog analysis demonstrates that early transcriptional responses to cold are more conserved in andropogoneae. Biorxiv doi: https://doi.org/10.1101/120303.

Zhang, Z., Xiao, J., Wu, J., Zhang, H., Liu, G., Wang, X., and Dai, L. (2012). Paraat: a parallel tool for constructing multiple protein-coding dna alignments. Biochemical and biophysical research communications **419**:779–781.

Zhu, Q., Cai, Z., Tang, Q., and Jin, W. (2016). Repetitive sequence analysis and karyotyping reveal different genome evolution and speciation of diploid and tetraploid tripsacum dactyloides. The Crop Journal **4**:247–255.

Figure legends

Figure 1. Morphological characteristics of inflorescences for representative species in Tripsacum and Zea. *Tripsacum dactyloides* (tripsacum, A-D) and *Zea mays* ssp. *parviglumis* (teosinte, E-H). (A) Early stage inflorescence. (B) Intermediate stage inflorescence with silks exerted from female spikelets. (C) magnified view of female spikelets. (D) mature inflorescence with both silks exerted from female spikelets (base) and anthers exerted from male spikelets (top). (E) Separate male and female inflorescences in teosinte. (F) magnified view of male inflorescence with anthers exerted. (G) magnified view of female inflorescence with silks exerted (H) hard fruitcases surrounding teosinte seeds (absent in domesticated maize). Bottom part present the latitude distribution of wild species in *Tripsacum* and *Zea* (outliers were removed), the order of which is consistent with that in Table 1.

Figure 2. Tripsacum Iso-seq data mapped onto the maize reference genome (RefGen_v3). (A) 10 chromosomes of maize. (B) Maize gene density in each chromosome. (C) Density of highly expressed maize genes (FPKM $>$ 4.15). (D) Tripsacum transcript density in each chromosome. (E) Density of syntenic gene pairs between maize and sorghum. (F) Density of sorghum-maize-tripsacum gene pairs. (G) Distribution of Long non-coding RNA (lncRNA) in tripsacum. (H) Distributions of homeologous genes pairs following the whole genome duplication which is shared by both tripsacum and maize.

Figure 3. Comparison of alternative splicing (AS) distribution in (A) different species using long- and short- reads (long reads data for maize taken from (Wang et al., 2016) and short reads data of maize and sorghum were from (Mei et al., 2017b)) and (B) subgenomes of maize and tripsacum. Identical shapes indicate genes with a conserved AS event. Solid line boxes mark cases most parsimoniously explained by a change in *trans*-regulation of AS, while dashed line boxes mark cases most parsimoniously explained by a change in *cis*-regulation of AS.

Figure 4. Species-specific substitution rates (A) Phylogenetic tree used in this analysis. Red, blue and black boxes respectively indicate target species, background species and outgroups. (B) Distribution of synonymous substitution rates (Ks) in orthologous gene groups across each target and background species. (C-E) Distribution of species-specific Ka/Ks ratios in five grasses.

Figure 5. Distribution of Ka/Ks ratios between maize and tripsacum. (A) Ratio of ratios Kernel density plot of the ratio of ratios between tripsacum and sorghum Ka/Ks values for genes identified as targets of selection between tropical and temperate maize lines and genes not identified as targets of selection in the same comparison. (B) Scatter plot showing the relationship between Ka/Ks ratios observed in maize (blue) and those observed in tripsacum (green) for genes having the highest Ka/Ks ratio in one of these two taxa. Orange triangles mark genes annotated as involved in glycerophospholipid metabolism, while red triangles mark genes involved in stress response.

Figure 6. Schematic plot of glycerophospholipid metabolism pathway containing the phospholipid synthesis processes. The enzyme names in red or blue that the protein coding sequences responsible for encoding these enzymes are experiencing elevated rates of protein sequence evolution in tripsacum or maize (respectively). Enzyme names in purple represent cases where different gene family members are experiencing accelerated rates of protein sequence evolution in each lineage. Each circle represents an individual gene which encodes an protein annotated as having the associated enzymatic activity. Circles are color

coded as described above, with circles in white representing genes which are not experiencing accelerated protein sequence evolution in either lineage. Pink boxes indicate final lipid head groups with full names spelled out, and the abbreviations used in this paper in parenthesis.

Figures

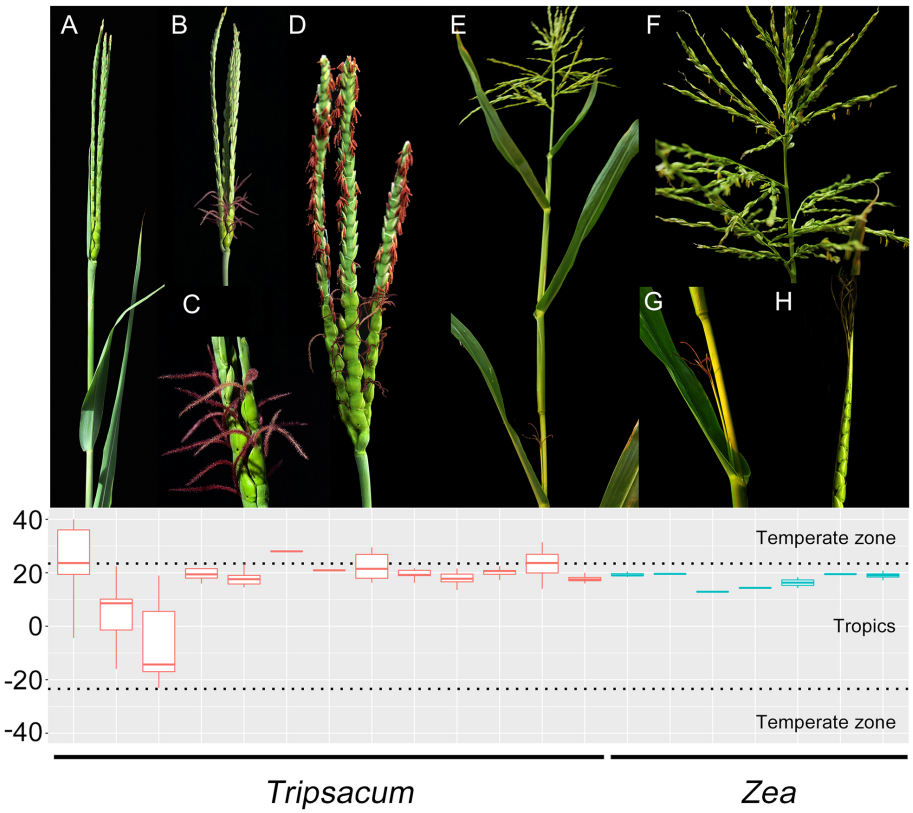

Figure 1. Morphological characteristics of inflorescences for representative species in Tripsacum and Zea. *Tripsacum dactyloides* (tripsacum, A-D) and *Zea mays* ssp. *parviglumis* (teosinte, E-H). (A) Early stage inflorescence. (B) Intermediate stage inflorescence with silks exerted from female spikelets. (C) magnified view of female spikelets. (D) mature inflorescence with both silks exerted from female spikelets (base) and anthers exerted from male spikelets (top). (E) Separate male and female inflorescences in teosinte. (F) magnified view of male inflorescence with anthers exerted. (G) magnified view of female inflorescence with silks exerted (H) hard fruitcases surrounding teosinte seeds (absent in domesticated maize). Bottom part present the latitude distribution of wild species in *Tripsacum* and *Zea* (outliers were removed), the order of which is consistent with that in Table 1.

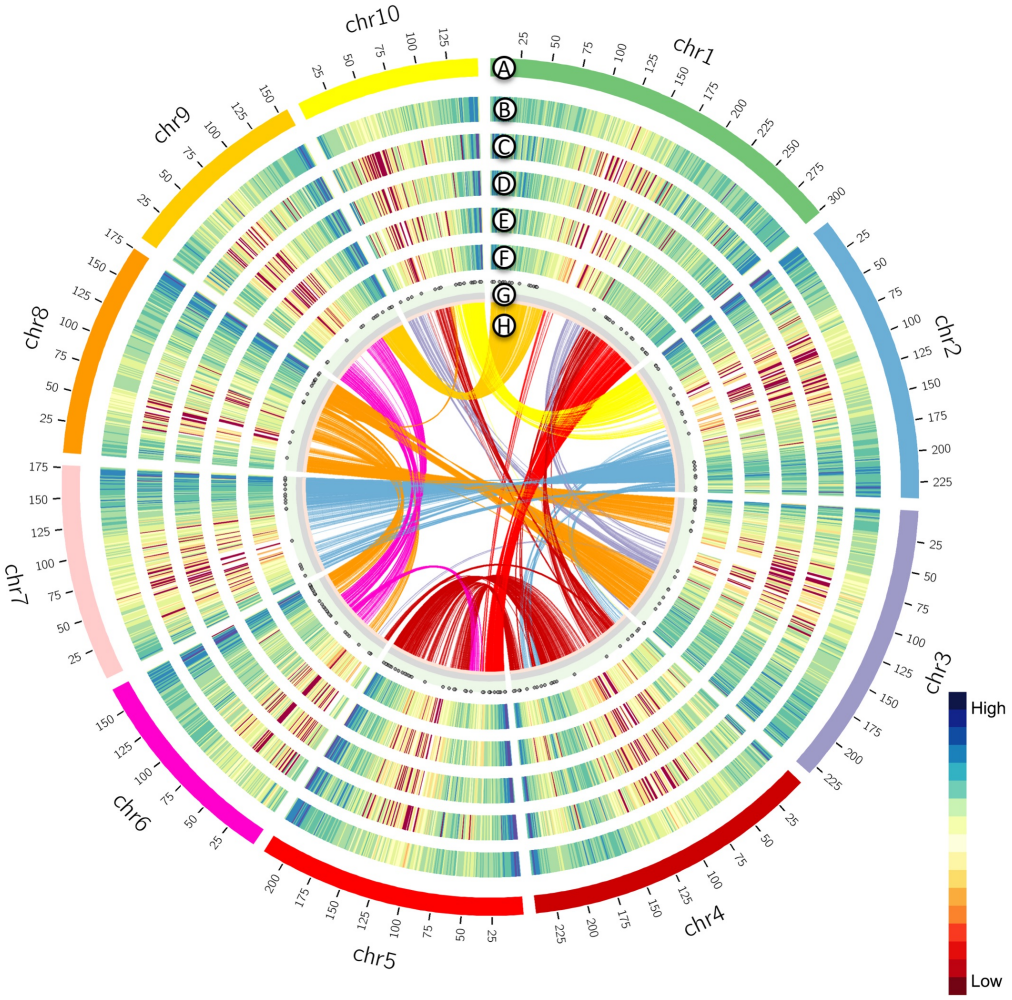

Figure 2. Tripsacum Iso-seq data mapped onto the maize reference genome (RefGen_v3). (A) 10 chromosomes of maize. (B) Maize gene density in each chromosome. (C) Density of highly expressed maize genes (FPKM $>$ 4.15). (D) Tripsacum transcript density in each chromosome. (E) Density of syntenic gene pairs between maize and sorghum. (F) Density of sorghum-maize-tripsacum gene pairs. (G) Distribution of Long non-coding RNA (lncRNA) in tripsacum. (H) Distributions of homeologous genes pairs following the whole genome duplication which is shared by both tripsacum and maize.

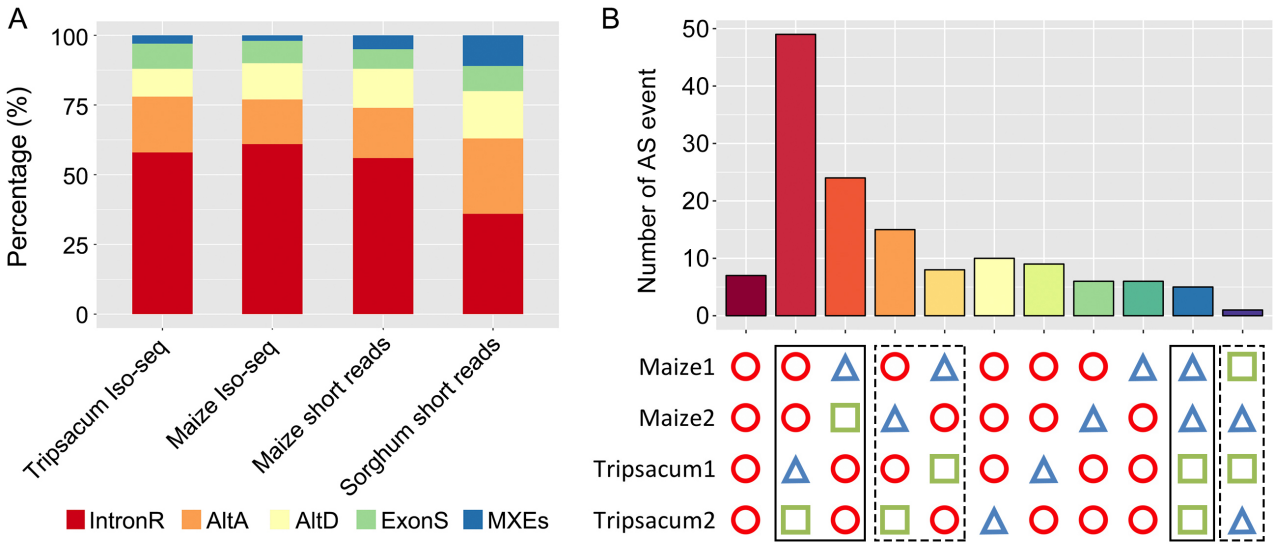

Figure 3. Comparison of alternative splicing (AS) distribution in (A) different species using long- and short- reads (long reads data for maize taken from (Wang et al., 2016) and short reads data of maize and sorghum were from (Mei et al., 2017b)) and (B) subgenomes of maize and tripsacum. Identical shapes indicate genes with a conserved AS event. Solid line boxes mark cases most parsimoniously explained by a change in *trans*-regulation of AS, while dashed line boxes mark cases most parsimoniously explained by a change in *cis*-regulation of AS.

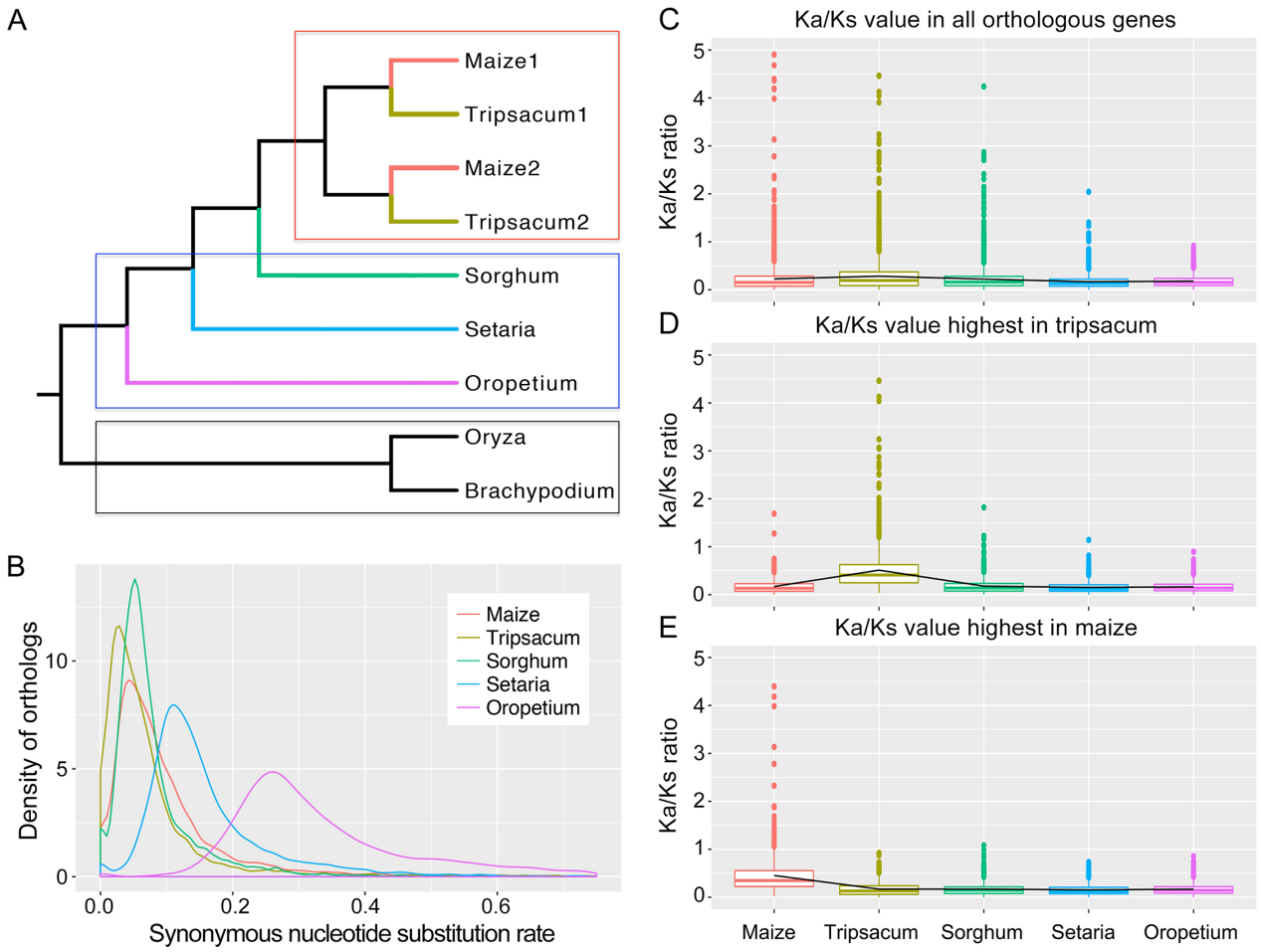

Figure 4. Species-specific substitution rates (A) Phylogenetic tree used in this analysis. Red, blue and black boxes respectively indicate target species, background species and outgroups. (B) Distribution of synonymous substitution rates (Ks) in orthologous gene groups across each target and background species. (C-E) Distribution of species-specific Ka/Ks ratios in five grasses.

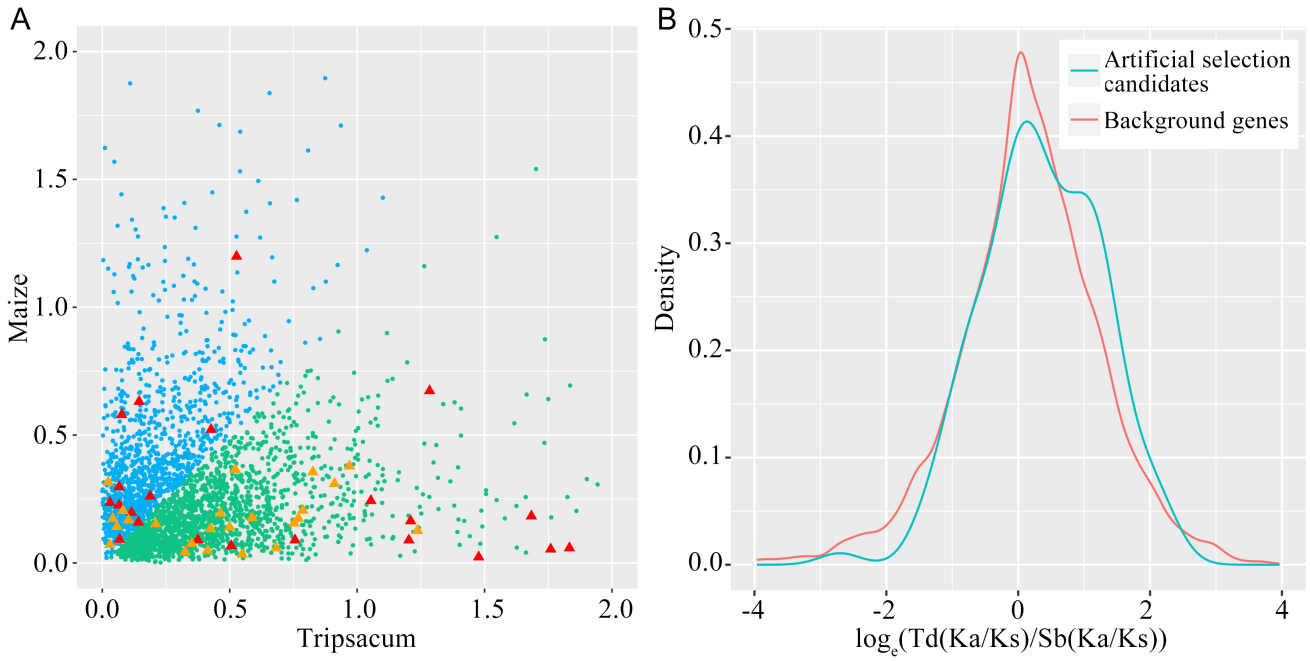

Figure 5. Distribution of Ka/Ks ratios between maize and tripsacum. (A) Ratio of ratios Kernel density plot of the ratio of ratios between tripsacum and sorghum Ka/Ks values for genes identified as targets of selection between tropical and temperate maize lines and genes not identified as targets of selection in the same comparison. (B) Scatter plot showing the relationship between Ka/Ks ratios observed in maize (blue) and those observed in tripsacum (green) for genes having the highest Ka/Ks ratio in one of these two taxa. Orange triangles mark genes annotated as involved in glycerophospholipid metabolism, while red triangles mark genes involved in stress response.

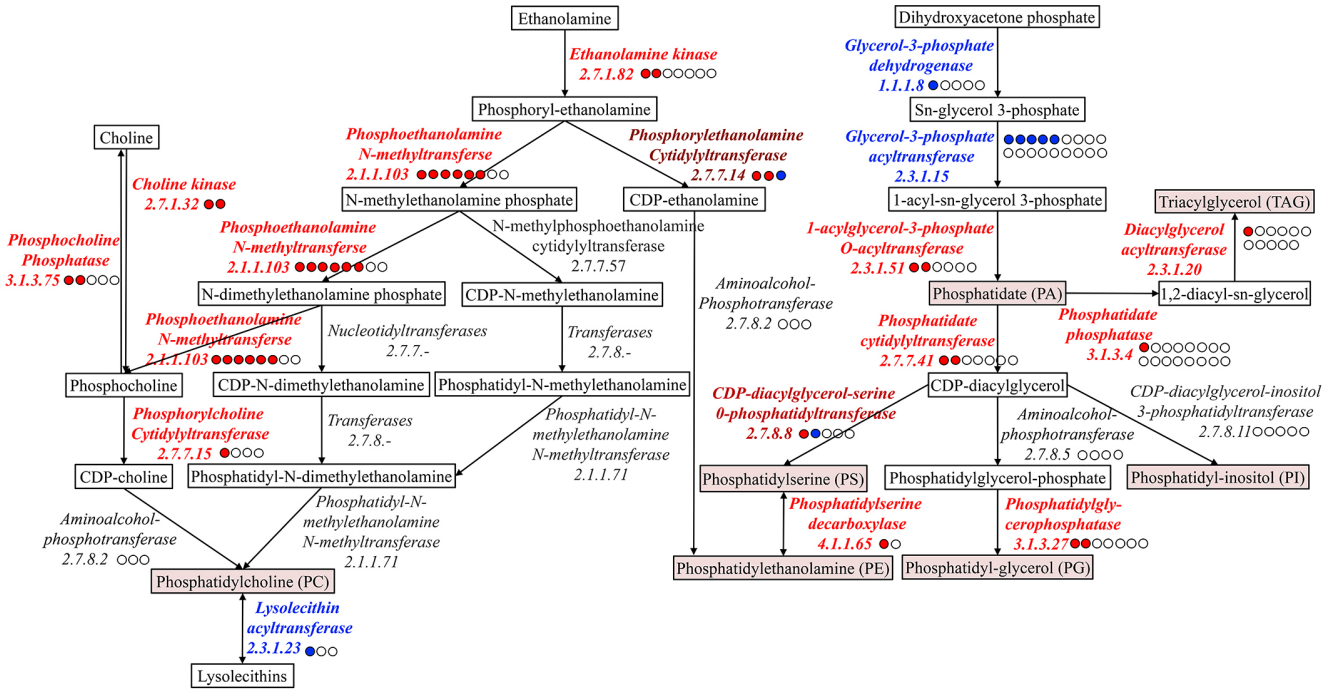

Figure 6. Schematic plot of glycerophospholipid metabolism pathway containing the phospholipid synthesis processes. The enzyme names in red or blue that the protein coding sequences responsible for encoding these enzymes are experiencing elevated rates of protein sequence evolution in tripsacum or maize (respectively). Enzyme names in purple represent cases where different gene family members are experiencing accelerated rates of protein sequence evolution in each lineage. Each circle represents an individual gene which encodes an protein annotated as having the associated enzymatic activity. Circles are color coded as described above, with circles in white representing genes which are not experiencing accelerated protein sequence evolution in either lineage. Pink boxes indicate final lipid head groups with full names spelled out, and the abbreviations used in this paper in parenthesis.

Tables

Table 1. Taxonomic comparison between genus *Tripsacum* and *Zea*.

Table 2. Patial list of conserved grass genes present in tripsacum but not in maize reference genome.

Table 3. Genes experiencing accelerate selections in tripsacum related to stress response.

Table 1. Taxonomic comparison between genus *Tripsacum* and *Zea*

Tripsacum[1]			*Zea*[2]		
Species	Life cycle	Ploidy	Species	Life cycle	Ploidy
T. dactyloides L. (Gamagrass)	perennial	diploid, tetraploid	*Z. perennis*	perennial	tetraploid
T. andersonii (Guatemala grass)	perennial	diploid, tetraploid	*Z. diploperennis*	perennial	diploid
T. australe	perennial	diploid	*Z. nicaraguensis*	annual	diploid
T. maizar	perennial	diploid	*Z. luxurians*	annual	diploid
T. laxum	perennial	diploid	*Z. m. huehuetenangensis*	annual	diploid
T. floridanum	perennial	diploid	*Z. m. mexicana*	annual	diploid
T. cundinamarce	perennial	diploid	*Z. m. parviglumis*	annual	diploid
T. zopilotense	perennial	diploid			
T. bravum	perennial	diploid			
T. latifolium	perennial	tetraploid			
T. pilosum	perennial	tetraploid			
T. lanceolatum	perennial	tetraploid			
T. peruvianum	perennial	tetraploid, pentaploid, hexaploid			

[1]References for *Tripsacum* are (Doebley, 1983; De Wet et al., 1981).

[2]References for *Zea* are (Iltis et al., 1979; Doebley and Iltis, 1980; Iltis and Doebley, 1980; Eubanks, 2001; Iltis and Benz, 2000).

Table 2. Patial list of conserved grass genes present in tripsacum but not in maize reference genome.

Seq in tripsacum	Ref-seq ID	Best hit species	Functions
c103397/f9p16/3344	ref|XM_004978753.1|	*Setaria italica*	inactive disease susceptibility protein LOV1
c142683/f8p4/3101	ref|XM_004962734.2|	*Setaria italica*	disease resistance RPP13-like protein 3
c105298/f4p4/3251	ref|XM_012843431.1|	*Setaria italica*	disease resistance protein RPM1-like
c134977/f1p2/3995	ref|XM_004979458.2|	*Setaria italica*	disease resistance protein RGA1
c71409/f4p4/2173	ref|XM_015789724.1|	*Oryza sativa*	disease resistance protein RPM1
c2068/f4p3/1285	ref|XM_015761770.1|	*Oryza sativa*	cysteine-rich receptor-like protein kinase
c93815/f3p1/3074	ref|XM_006659166.2|	*Oryza brachyantha*	S-receptor-like serine/threonine-protein kinase RLK1
c24204/f1p3/1320	ref|XM_003558788.3|	*Brachypodium distachyon*	ABC transporter C family member 13
c73653/f9p17/1867	ref|XM_003579206.3|	*Brachypodium distachyon*	LRR receptor-like serine/threonine-protein kinase
c99674/f2p3/3370	ref|XM_020329183.1|	*Aegilops tauschii*	receptor kinase-like protein Xa21
c140099/f1p27/1232	ref|XM_020305102.1|	*Aegilops tauschii*	myrosinase-binding protein 2-like

Table 3. Genes experiencing accelerate selections in tripsacum related to stress response.

Tripsacum ID	Orthologs in maize	Ka/Ks in tripsacum	Ka/Ks in maize	Functions
Td.547	GRMZM2G010000	1.83296	0.05825	Heat shock protein
Td.8489	GRMZM2G101287	1.7595	0.05365	Universal stress protein domain containing protein
Td.8643	GRMZM2G069099	1.68381	0.1831	Cold-induced protein subunit 5
Td.594	GRMZM2G020940	1.47686	0.02288	Stress responsive protein
Td.7706	AC207656.3_FG002	1.28433	0.6731	Auxin response factor 16
Td.5834	GRMZM2G021194	1.20983	0.16403	ERD (early-responsive to dehydration stress)
Td.11525	GRMZM2G150367	1.20303	0.08928	Stress-related protein
Td.2246	GRMZM2G103909	1.05386	0.24332	Drought-responsive family protein
Td.11786	GRMZM2G168552	0.75617	0.08919	Abscisic stress-ripening
Td.7779	GRMZM2G118884	0.50582	0.06662	Salt tolerance homolog2

Supplemental Information (SI)

Table S1. Summary of sequence data produced.

Table S2. Number of maize genes with one or more identified tripsacum isoforms.

Figure S1. Workflow of Iso-Seq bioinformaics analysis for tripsacum using the SMRT-Analysis software package (SMRT Pipe v2.3.0).

Figure S2. Comparison of length distribution of lncRNAs identified using Pacbio sequencing data between maize and tripsacum.

Figure S3. Example of one tripsacum transcript spans two maize gene models but correlated with one sorghum gene model through GEvo analysis.

Figure S4. Proportions of alternative splicing (AS) types (Intron retention, IntronR; Exon skipping, ExonS; Alternative donor, AltD; Alternative acceptor, AltA; Mutually exclusive exons, MXEs; Complicated AS, CompAS; More complicated AS, MoreCompAS) found in (A) tripsacum and (B) maize using PacBio long sequences. (C) Proportion of conserved AS types between maize and tripsacum.

Figure S5. Ks distribution of orthologous gene pairs between each two of species pair-wisely (bin size = 0.05). (A) divergence between tripsacum and maize (td-zm), maize and sorghum (zm-sb), sorghum and setaria (sb-si), setaria and oropetium (si-or), their divergence were shown by the peak of each pair. (B) divergence of maize with other species. (C) divergence of tripsacum with other species. (D) divergence of sorghum with other species. (E) divergence of setaria with other species. (F) divergence of oropetium with other species.

Figure S6. Distribution of species-specific Ka/Ks ratios including the homeologous gene quartets in both tripsacum and maize shared the WGD. (A) distribution of Ka/Ks ratios in orthologous genes sets. (B) increased Ka/Ks ratios in maize1. (C) increased Ka/Ks ratios in maize2. (D) increased Ka/Ks ratios in tripsacum1. (E) increased Ka/Ks ratios in tripsacum2.

Figure S7. (A) Ka/Ks distributions in tripsacum between lipid genes and other functional genes. (B-F) Phenotype changes between tripsacum (left) and maize (right) in continuous five days after cold treatment at 4 °C for two days.

Supplemental Tables

Table S1. Summary of sequence data produced.

Samples	Cell	Total ROIs[1]	Total FL Reads[2]	Total FLNC Reads[3]	Average FLNC Length	FLNC
1-2 kb	2	151,126	79,733	79,340	1,364 bp	52.5%
2-3 kb	2	181,154	94,117	93,694	2,376 bp	51.7%
3-6 kb	2	199,791	95,071	94,152	3,323 bp	47.1%
Total	6	532,071	268,921	267,186	2,354 bp	50.2%

[1]ROIs: Reads Of Inserts;

[2]FL reads: full length reads, defined by the presence of both the 5' and 3' primer and barcode.

[3]FLNC Reads: full-length and non-chimeric reads.

Table S2. Number of maize genes with one or more identified tripsacum isoforms.

Isoforms per gene	Number of genes		Isoforms per gene	Number of genes
1	7,633		14	4
2	2,880		15	5
3	1,280		16	2
4	580		17	1
5	283		18	1
6	164		19	2
7	94		24	1
8	47		29	1
9	44		33	1
10	26		43	1
11	19		50	1
12	9		55	1
13	8		83	1

Supplemental Figures

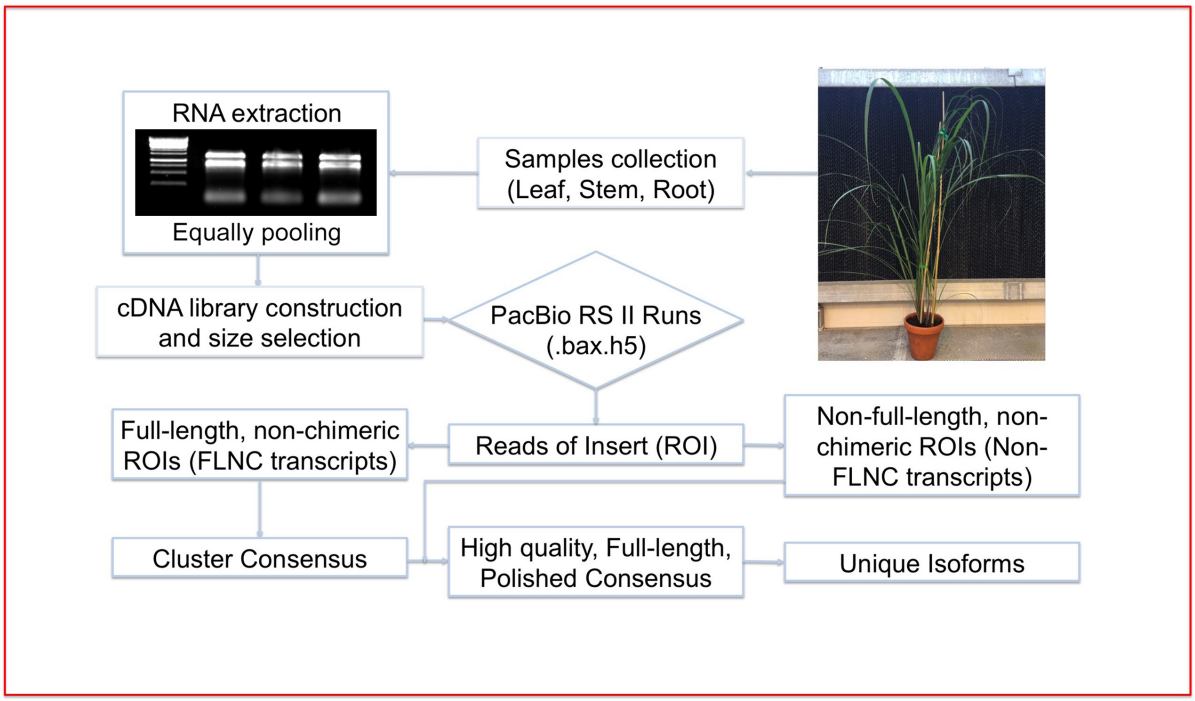

Figure S1. Workflow of Iso-Seq bioinformaics analysis for tripsacum using the SMRT-Analysis software package (SMRT Pipe v2.3.0).

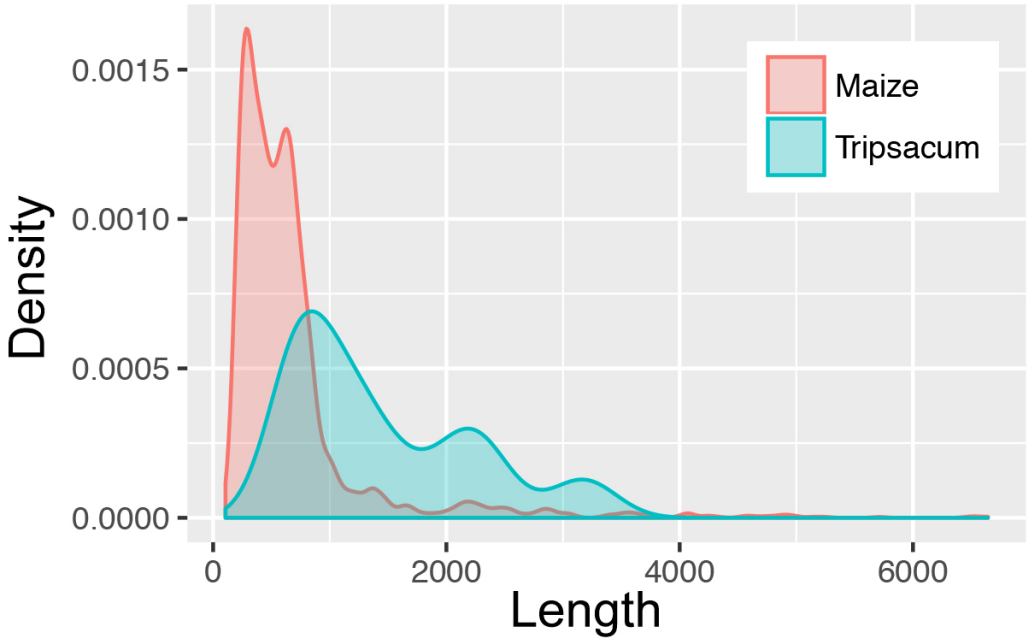

Figure S2. Comparison of length distribution of lncRNAs identified using Pacbio sequencing data between maize and tripsacum.

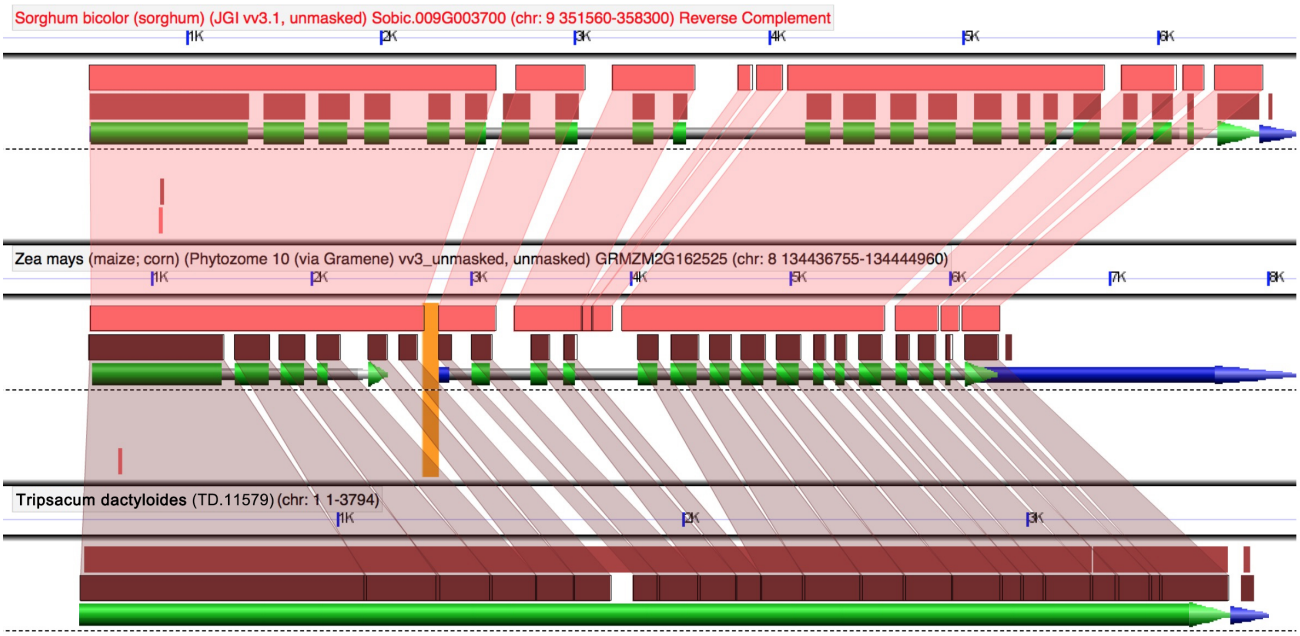

Figure S3. Example of one tripsacum transcript spans two maize gene models but correlated with one sorghum gene model through GEvo analysis.

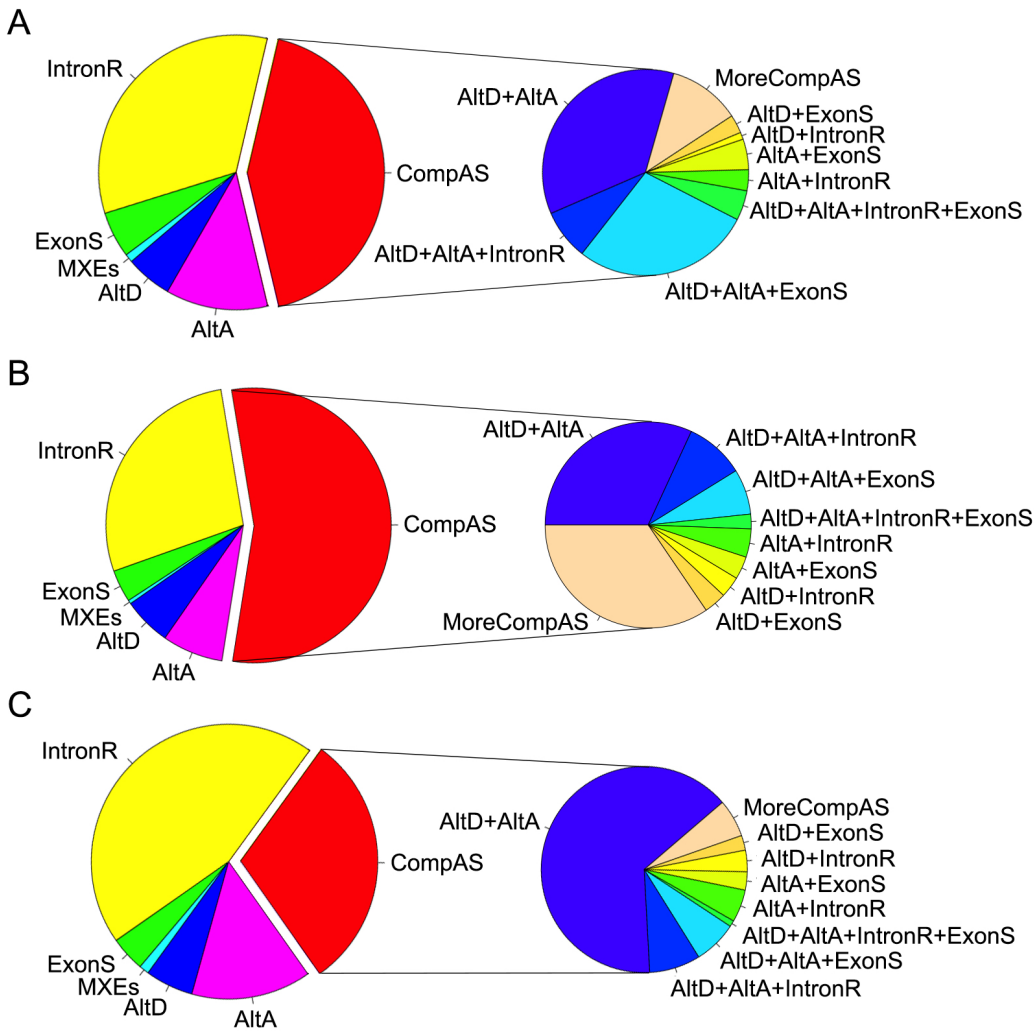

Figure S4. Proportions of alternative splicing (AS) types (Intron retention, IntronR; Exon skipping, ExonS; Alternative donor, AltD; Alternative acceptor, AltA; Mutually exclusive exons, MXEs; Complicated AS, CompAS; More complicated AS, MoreCompAS) found in (A) tripsacum and (B) maize using PacBio long sequences. (C) Proportion of conserved AS types between maize and tripsacum.

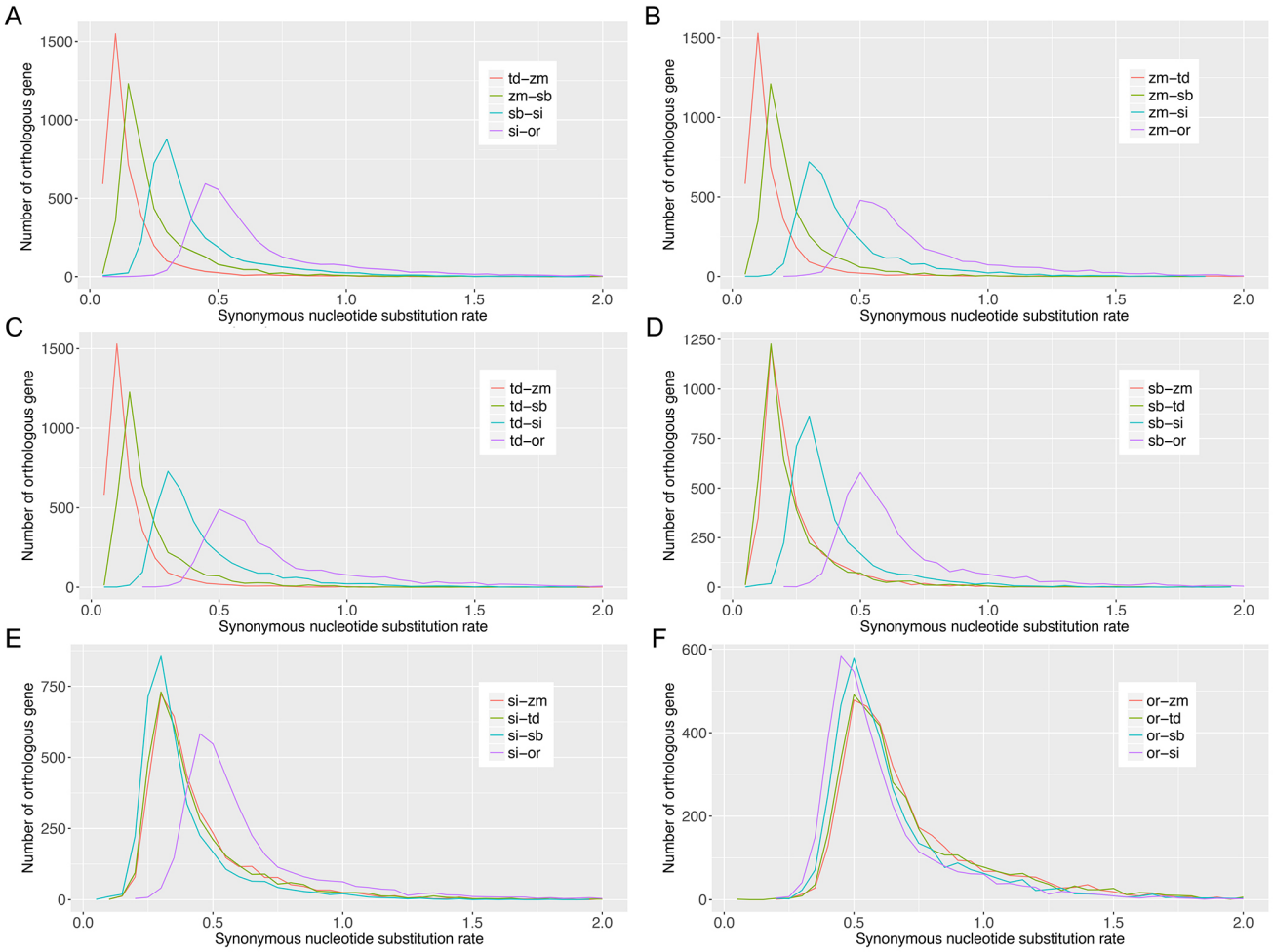

Figure S5. Ks distribution of orthologous gene pairs between each two of species pair-wisely (bin size = 0.05). (A) divergence between tripsacum and maize (td-zm), maize and sorghum (zm-sb), sorghum and setaria (sb-si), setaria and oropetium (si-or), their divergence were shown by the peak of each pair. (B) divergence of maize with other species. (C) divergence of tripsacum with other species. (D) divergence of sorghum with other species. (E) divergence of setaria with other species. (F) divergence of oropetium with other species.

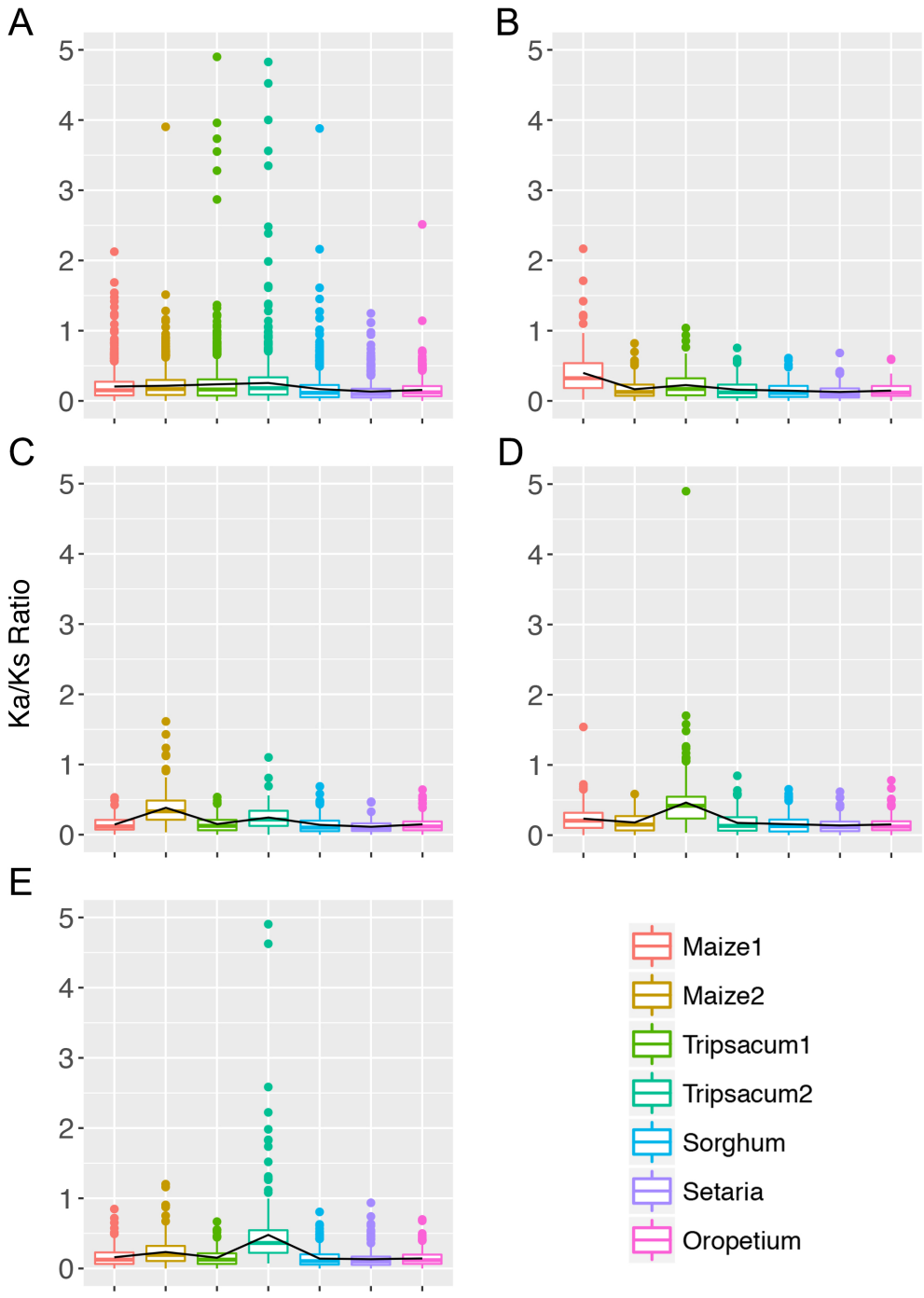

Figure S6. Distribution of species-specific Ka/Ks ratios including the homeologous gene quartets in both tripsacum and maize shared the WGD. (A) distribution of Ka/Ks ratios in orthologous genes sets. (B) increased Ka/Ks ratios in maize1. (C) increased Ka/Ks ratios in maize2. (D) increased Ka/Ks ratios in tripsacum1. (E) increased Ka/Ks ratios in tripsacum2.

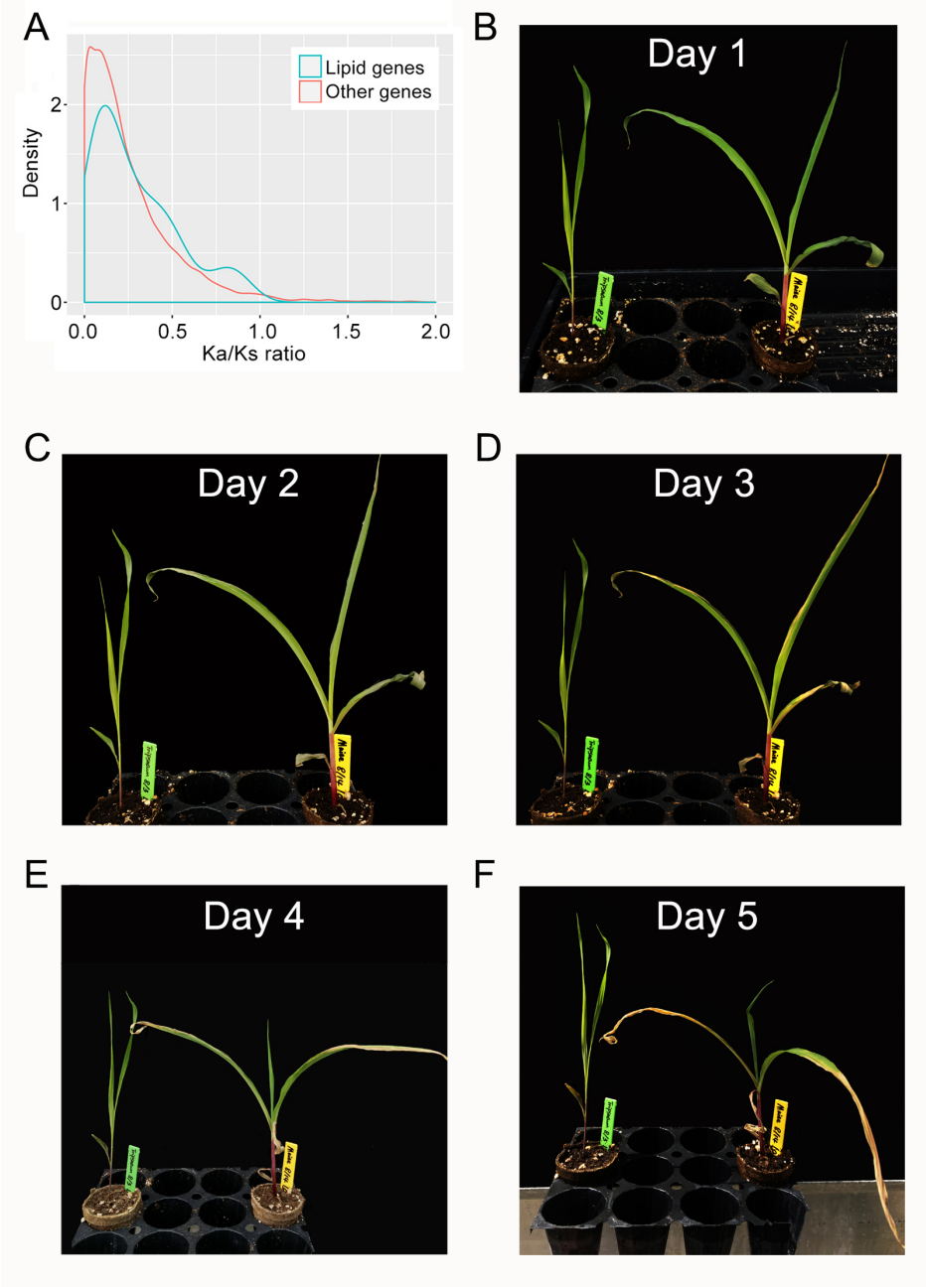

Figure S7. (A) Ka/Ks distributions in tripsacum between lipid genes and other functional genes. (B-F) Phenotype changes between tripsacum (left) and maize (right) in continuous five days after cold treatment at 4 °C for two days.