

Diploids Derived from Polyploids: Genetic Characteristics of Four Novel Interspecific *Sorghum* Populations

Wenqian Kong^{1,3}, Pheonah Nabukalu², Stan Cox², Jaxk Reeves³ and Andrew H. Paterson¹

¹Plant Genome Mapping Laboratory, Institute of Plant Breeding, Genetics & Genomics (IPBGG), University of Georgia, Athens, GA, USA

²The Land Institute, Salina, Kansas, USA

³Department of Statistics, University of Georgia, Athens, GA, USA

¹Corresponding author: Andrew H. Paterson, patersonah@gmail.com

Abstract

Polyploidy has repeatedly influenced the evolution of grasses, yet direct observations of how polyploid-derived chromosomes behave when returned to diploidy are exceedingly rare. Interspecific crosses between diploid *Sorghum bicolor* and tetraploid hybrids derived from *S. halepense* generate mixed-ploidy progeny, enabling analysis of chromosome transmission immediately following polyploid formation. Using genome-wide SNP markers, we characterized chromosomal inheritance in two diploid and two tetraploid families derived from these crosses. Genotype-dosage profiles alone distinguished diploids from tetraploids with complete accuracy, reflecting sharp ploidy-dependent differences in dosage-class distributions. Although diploid progeny retained most halepense-derived haplotypes, they exhibited several genomic intervals in which *S. bicolor* alleles formed long, non-random runs of homozygosity that remained polymorphic in their tetraploid counterparts. These patterns, together with reciprocal segregation distortion across families, indicate that the transition from tetraploidy to diploidy exposes allelic combinations that differ markedly in viability or transmission. Analyses of flowering time further revealed that diploid and tetraploid derivatives possess distinct genetic architectures, with major loci mapping to different chromosomal regions across ploidy levels. Collectively, these results show that early diploidization involves structured retention and loss of parental haplotypes shaped by both selective and structural constraints. The diploid extractions characterized here provide a rare empirical system for dissecting the initial phases of diploidization and a practical route for mobilizing polyploid-derived variation into sorghum improvement.

Keywords: polyploidy; diploidization; *Sorghum*; introgression; segregation distortion; QTL mapping

Introduction

Polyploidy has played a central role in angiosperm evolution, repeatedly generating duplicated chromosome sets that introduce new opportunities—and constraints—for genome function. Newly formed polyploids often exhibit instability as duplicated loci, homoeologous chromosomes, and dosage relationships interact in ways not encountered in diploids. Polyploid lineages that persist must therefore stabilize the behavior of multiple chromosome complements within a shared nucleus. This stabilization typically involves coordinated gene loss, chromosomal restructuring, and reductions in chromosome number that collectively drive the return to a more diploid-like state. Such diploidization processes can influence the ecological range, competitive ability, and evolutionary potential of polyploids, helping explain their disproportionate representation among invasive plant taxa (Pandit *et al.* 2006).

Polyploids also frequently differ from their diploid relatives in morphology, physiology, and genomic composition (Parisod *et al.* 2010). A well-studied example in the grasses is *Sorghum halepense* (johnsongrass), an allotetraploid ($2n = 4x = 40$) formed through hybridization between diploid *S. bicolor* and *S. propinquum*. The species combines vigorous rhizomatous growth, recurrent flowering, and broad ecological amplitude, and it hybridizes readily with cultivated sorghum, contributing to persistent crop–weed gene

flow (Arriola and Ellstrand 1996, 1997; Morrell *et al.* 2005). These features illustrate how polyploid genomes can integrate genetic material from divergent parents into combinations that enhance both adaptability and invasiveness.

Despite its reputation as a weed, *S. halepense* harbors alleles of considerable value for sorghum improvement. Traits such as perenniality, rhizome development, and abiotic-stress tolerance have motivated efforts to introgress halepense-derived segments into *S. bicolor*. Early work at The Land Institute showed that interspecific tetraploid populations produced via unreduced-gamete formation capture extensive phenotypic and genomic variation relevant to perennial grain development (Piper and Kulakow 1994; Cox *et al.* 2018a; Kong 2017). Yet tetraploid genomes introduce analytical challenges: polysomic inheritance obscures segregation patterns, recombination among homoeologous chromosomes complicates map construction, and introgressed segments may be difficult to trace or transfer into diploid breeding lines.

A striking observation from these breeding efforts was that certain tetraploid intermediates occasionally produce viable diploid offspring when crossed to diploid *S. bicolor* (Cox *et al.* 2018b). These “diploid extractions” effectively return polyploid-derived chromosomal segments to a diploid genomic context. Such events create an unusually tractable system for studying the early stages of diploidization by allowing direct comparison of homologous chromosomal regions maintained in both diploid and tetraploid backgrounds. Systems that capture this transition from polyploid to

diploid inheritance with shared parental genomes are exceedingly rare in plants.

In this study, we characterize two diploid and two tetraploid interspecific families derived from crosses between diploid *S. bicolor* and *S. halepense*-derived tetraploids. Using genome-wide SNP markers, we (i) quantify patterns of parental haplotype transmission, (ii) identify genomic regions exhibiting biased introgression or segregation distortion, (iii) compare the retention and organization of *S. halepense*- versus *S. bicolor*-derived segments across ploidy levels, and (iv) resolve the genetic architecture of flowering time in both diploid and tetraploid derivatives. By jointly analyzing diploid extractions and their tetraploid progenitors, we gain insight into chromosome behavior as polyploid-derived alleles are returned to diploidy, advancing understanding of early diploidization processes while identifying polyploid-origin alleles with potential value for sorghum improvement.

Materials and Methods

Population development

We analyzed two related interspecific families, each comprising paired diploid and tetraploid populations derived from crosses between cultivated diploid *Sorghum bicolor* and *Sorghum halepense*-derived tetraploids. The relationships among parental lines and the segregation of mixed-ploidy F_1 progeny into F_2 diploid and tetraploid families are schematized in Figure 1.

The S2039 families trace to a cross between diploid *S. bicolor* K102 ($2n = 20$) and tetraploid T254>123 (hereafter T254; $2n = 40$), an intermediate developed by recurrently backcrossing *S. halepense* to *S. bicolor*. Segregation of ploidy in the F_1 generation yielded a diploid F_2 family (S2039D; $N = 180$) and a tetraploid F_2 family (S2039T; $N = 122$).

Similarly, the S2172 families were obtained by crossing diploid *S. bicolor* K105 with tetraploid T115>B29A (hereafter T115), another intermediate generated by repeated introgression of *S. halepense* into a cultivated sorghum background. The resulting F_2 progeny segregated into 300 diploid individuals (S2172D) and 62 tetraploid individuals (S2172T). Details of tetraploid intermediate development and recovery of diploid progeny from interpollid crosses are described in Cox et al. (2018b).

Genotyping-by-sequencing

Leaf samples from all individuals were flash-frozen at -80°C and lyophilized for 48 h prior to DNA extraction. Genomic DNA was isolated using the rapid, polysaccharide- and polyphenol-free protocol of Aljanabi et al. (1999). Genotyping-by-sequencing (GBS) libraries were constructed following the two-enzyme PstI-MspI protocol of Poland et al. (2012). DNA concentrations were normalized to $20\text{ ng }\mu\text{L}^{-1}$ before library preparation. Amplified libraries were size-selected to 250–300 bp using the QIAquick Gel Extraction Kit (QIAGEN).

Libraries were sequenced on an Illumina NextSeq platform (RRID:SCR_016400) with 150 bp single-end reads. Sequence reads were aligned to the *S. bicolor* BTx623 v3.1.1 reference genome (McCormick et al. 2018) using the Burrows–Wheeler Aligner (BWA; RRID:SCR_010910) (Li and Durbin 2009). Variant discovery and initial quality filtering were performed with the Genome Analysis Toolkit (GATK; RRID:SCR_001876) (McKenna et al. 2010).

Phenotyping for flowering time

The four populations (S2039D, S2039T, S2172D, and S2172T) were grown at The Land Institute (Salina, Kansas, USA) in 2018 and

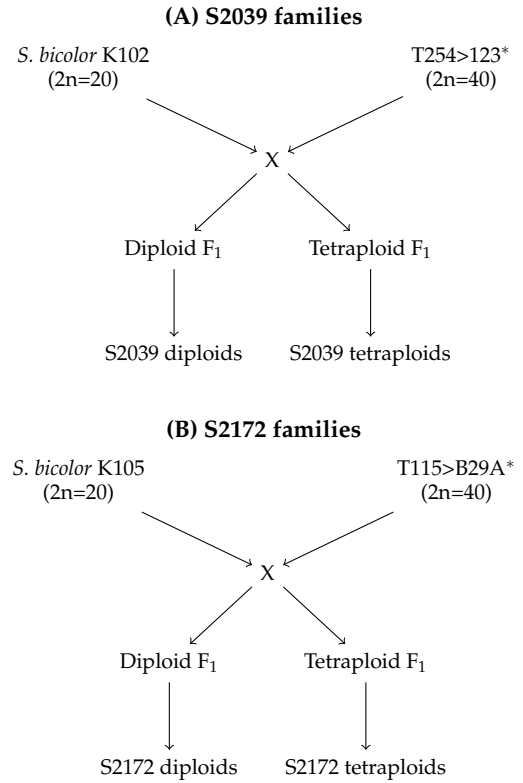


Figure 1 Parental and progeny relationships for the S2039 and S2172 diploid and tetraploid families. Diploid *S. bicolor* lines K102 and K105 were crossed to tetraploid intermediates T254>123 and T115>B29A, respectively. Mixed-ploidy F_1 progeny segregated into diploid and tetraploid classes, which were advanced to generate the S2039 and S2172 families. *Tetraploid intermediates were produced by recurrent backcrossing of *S. bicolor* to *S. halepense*.

2019 using a randomized design with two blocks each year. In 2018, unfavorable weather necessitated two planting dates (5 and 8 June), and planting date was modeled as a block effect. In 2019, all entries were sown on 20 May.

Flowering time (FT) was recorded for each row as the number of days from planting until approximately 50% of plants in that row had flowered. After quality control for both genotypes and phenotypes, complete data were available for 179 individuals in S2039D, 81 in S2039T, 200 in S2172D, and 53 in S2172T.

Parental analysis and run-length methodology

To characterize introgression patterns in the tetraploid intermediates, we retained 13,991 SNPs segregating across five genotypes: K105, two *S. halepense* accessions (G9N and G2012), and the tetraploids T115 and T254. Sites polymorphic in K105 relative to the BTx623 reference were excluded so that deviations from the reference would primarily reflect *S. halepense*-derived variation.

For each SNP, we defined a composite indicator of *S. halepense* ancestry, G_{combined} , coded as 0 if both G9N and G2012 matched the BTx623 reference, and 1 if at least one of the two *halepense* accessions carried an alternative allele. Genotypes in T115 and T254 were coded as 0 (homozygous reference), 1 (heterozygous), or 2 (homozygous alternative). Within each tetraploid, we then scanned along chromosomes for consecutive markers sharing the same genotype class, treating these as “runs” of 0, 1, or 2.

Following Kong et al. (2020), the minimum run length k_g considered non-random for genotype class g was defined by

$$1 - \left[\left(\frac{n_g}{N} \right)^{k_g} (N + 1 - k_g) \right] \geq 1 - \alpha,$$

where N is the total number of markers and n_g is the number of markers in class g across the genome, with $\alpha = 0.05$. Runs exceeding this threshold were interpreted as clusters of unusually long homozygous or heterozygous segments, consistent with introgressed chromosomal regions.

SNP calling and filtering in S2039 and S2172

For joint analysis of diploid and tetraploid progeny, variants were called across all individuals under a tetraploid genotype model to facilitate comparisons based on dosage. Only biallelic SNPs with minor allele frequency (MAF) greater than 0.05 were retained.

Depth filters were applied to reduce spurious homozygous calls. Specifically, homozygous genotypes with read depth < 6 were masked, because the probability of observing at least one alternative allele under a true heterozygote at depth 6 is

$$1 - 2 \times 0.5^6 = 0.9688,$$

indicating that low-depth homozygous calls are unlikely to reflect genuine heterozygotes. Markers with more than 20% missing data across individuals were removed. To minimize redundancy, markers in near-perfect linkage ($r > 0.99$) were identified, and one representative per highly correlated set was retained. After filtering, 10,102 SNPs for S2039 and 5,945 SNPs for S2172 remained, and genotypes were expressed as tetraploid dosage classes (0–4).

Linear discriminant analysis for ploidy classification

To distinguish diploid from tetraploid individuals using genotype data, we summarized each individual by a five-element composition vector giving the proportion of SNPs in dosage classes 0, 1, 2, 3, and 4. These vectors were used as predictors in a linear discriminant analysis (LDA), with cytologically verified diploid or tetraploid status as the response, to evaluate how well dosage profiles alone separate the two ploidy classes.

Multidimensional scaling of genomic relationships

For each family, we constructed a combined genotype matrix spanning both diploid and tetraploid individuals, using dosage-based calls and removing markers with $r > 0.99$ to reduce redundancy. For distance-based analyses, genotypes were recoded into three classes: 0 (homozygous reference), 1 (heterozygous; dosage 1–3), and 2 (homozygous alternative). Pairwise genomic distances between individuals i and j were then computed using a modified Hamming distance:

$$d_{ij} = \frac{1}{2m'} \sum_{k=1}^m |x_{ik} - x_{jk}|,$$

where m' is the number of markers scored in both individuals (i.e., non-missing in i and j), and x_{ik} and x_{jk} denote recoded genotypes at marker k . Classical multidimensional scaling (MDS) of the resulting distance matrix was performed using the `cmdscale` function in R (RRID:SCR_001905).

Comparative genomic composition of diploid and tetraploid progeny

To compare genomic composition between diploid and tetraploid progeny within each family, we estimated the alternative-allele frequency (p) for every SNP in each population and assigned it to one of three classes: reference-enriched ($p < 0.25$; $g = 0$), approximately balanced contributions ($0.25 \leq p \leq 0.75$; $g = 1$), or alternative-enriched ($p > 0.75$; $g = 2$). As above, we identified runs of consecutive markers sharing the same class and applied the run-length thresholds derived from the global frequency of each class to identify unusually long segments. The distribution of markers among these three categories in each population is summarized in Table 1.

Genetic map construction

For QTL mapping in diploid families, we applied additional quality filters tailored to linkage analysis. GATK hard-filter criteria were used at the variant level:

QD < 2.0 || FS > 60.0 || MQ < 40.0 ||
MQRankSum < -12.5 || ReadPosRankSum < -8.0

SNPs with MAF ≤ 0.05 or more than 50% missing data were discarded. The final mapping datasets comprised 5,684 SNPs for S2039D and 4,697 SNPs for S2172D. Linkage groups and initial recombination fractions were estimated in `onemap` (Margarido et al. 2007), and final inter-marker distances were refined using `R/qt1` (Broman et al. 2003) (RRID:SCR_001821).

QTL mapping for flowering time

Quantitative trait loci (QTL) for flowering time in the diploid populations were mapped by interval mapping (Lander and Botstein 1989) implemented in `R/qt1`. Genome-wide significance thresholds for the LOD score were obtained by permutation testing (1,000 permutations; threshold LOD = 3.5). Significant QTL identified in initial scans were incorporated as additive covariates in subsequent models, and final multiple-QTL models were used to estimate the proportion of phenotypic variance explained by each locus.

Sample sizes in tetraploid families were insufficient for conventional linkage mapping. For these populations, we therefore conducted single-marker association tests with 2018 flowering-time data to identify markers showing evidence of association, recognizing that these results are exploratory and provided primarily for comparison with the diploid QTL.

Statistical analysis

All analyses were performed in R (RRID:SCR_001905). Variant calling and hard filtering followed GATK best-practice recommendations (RRID:SCR_001876). Linkage maps were constructed using `onemap` and `R/qt1` (Margarido et al. 2007; Broman et al. 2003) with a genotyping error rate of 0.005. Genome-wide permutation tests (1,000 iterations) were used to derive LOD thresholds for QTL detection. For tetraploid single-marker tests, we report raw p -values; full marker-wise outputs are provided in Supplementary Data, enabling readers to apply alternative multiple-testing corrections if desired. The phenotypic design included two blocks per year, and sample sizes for all trait analyses are summarized in Table S1.

Results and Discussion

Parental genomic composition and similarity to *S. halepense*

Summary statistics in Table 2 indicate that both tetraploid intermediates, T254 and T115, retain substantial *S. halepense* ancestry while also carrying discrete tracts of alleles that match *S. bicolor*. For sites where the composite *S. halepense* indicator G_{combined} (based on accessions G9N and G2012) matches the *S. bicolor* reference (i.e., $G_{\text{combined}} = 0$), 16.0% of genotypes in T254 and 14.4% in T115 are neither reference-like nor halepense-like. These “unknown” alleles likely represent additional *S. halepense* haplotypes not sampled by G9N or G2012, novel mutations, or *S. bicolor* polymorphisms absent from the BTx623 reference.

For sites where G_{combined} carries at least one alternative allele ($G_{\text{combined}} = 1$), the two tetraploids differ in the relative contributions of reference versus halepense-like haplotypes. At these loci, 16.5% (T254) and 28.8% (T115) of alleles match the *S. bicolor* reference, whereas 60.6% (T254) and 47.9% (T115) align with the halepense-derived haplotypes from G9N or G2012 (Table 2). On balance, both intermediates are more similar to *S. halepense* than to *S. bicolor*, but they contain sizeable and evidently non-random blocks of *S. bicolor* introgression.

Clustering of parental genotypes based on genome-wide SNP data (Figure 2) is consistent with these allele-frequency patterns. T115 and T254 group more tightly with the halepense accessions than with cultivated *S. bicolor* K105, reinforcing the view that recurrent introgression has not erased the fundamentally halepense-like character of the tetraploid genomes.

Run-length analysis further resolves how *S. bicolor*-like tracts are distributed along chromosomes in the tetraploid parents. In T115, eight regions exceed the threshold for unusually long runs of homozygous *S. bicolor*-like alleles (minimum run length $k = 13$, $p < 0.05$; Figure S1), whereas T254 carries two such regions (Figure S2). Several of these tracts coincide with a region on chromosome 5 that includes the rDNA locus previously reported to be depleted for *S. halepense* alleles (e.g., Kong et al. 2020). Recurrence of this pattern across independent interspecific backgrounds suggests that a particular *S. bicolor* haplotype in the rDNA region has been consistently favored during the development of halepense-derived tetraploids.

Diploid genetic maps and segregation distortion

The S2039D map was constructed from 180 F₂ individuals genotyped at 504 SNP markers (Figure 3). The map contains an average of 50.4 markers per chromosome, with chromosome 3 carrying the most markers (64) and chromosome 8 the fewest (36). Total map length is 2,095.6 cM, corresponding to a mean intermarker distance of 4.28 cM; the average largest gap per chromosome is 12.99 cM (Table 3).

Across the genome, genotype frequencies in S2039D are 27.2% AA (homozygous reference), 47.9% AB (heterozygous), and 24.9% BB (homozygous alternative), close to the 1:2:1 expectation for a fully informative F₂, but with clear local departures. Chromosome 4 is the most *S. bicolor*-biased (32.3% AA, 47.7% AB, 19.9% BB), whereas chromosome 6 is the most *S. halepense*-biased (16.8% AA, 48.5% AB, 34.7% BB). Extended segments of segregation distortion ($p < 0.05$) occur on chromosomes 2 and 6 (Figure 4); distorted loci on chromosome 2 preferentially transmit *S. bicolor* alleles, while those on chromosome 6 favor *S. halepense* alleles.

For S2172D, 300 F₂ individuals were genotyped at 610 SNPs (Figure 5). This map averages 61.0 markers per chromosome, with

chromosome 3 carrying 90 markers and chromosome 7 carrying 38. The total map length is 2,671.1 cM, and mean intermarker spacing is 4.49 cM; the largest single gap is 37.0 cM on chromosome 1 (Table 3). Genome-wide genotype frequencies are 27.3% AA, 46.8% AB, and 25.9% BB, again similar to Mendelian expectations at the global scale.

Patterns of allele enrichment across chromosomes in S2172D broadly parallel those in S2039D. Chromosome 4 is again the most *S. bicolor*-enriched (30.9% AA, 46.7% AB, 22.4% BB), whereas chromosomes 8 (22.2% AA, 45.2% AB, 32.5% BB) and 6 (22.3% AA, 46.8% AB, 30.9% BB) are the most *S. halepense*-enriched. Segregation distortion is more diffusely distributed than in S2039D, likely reflecting higher power from the larger sample size. Chromosomes 6 and 10 each harbor extended series of distorted markers (> 6 consecutive markers with $p < 0.05$; Figure 6), and as in S2039D, distorted loci on chromosome 6 preferentially transmit *S. halepense* alleles.

The distorted regions on chromosome 6 in the two families share a broad physical interval. In S2039D, significantly distorted markers span 47.76–59.34 Mb, and in S2172D they span 51.7–58.83 Mb. The overlap suggests recurrent transmission bias or selection acting on the same genomic region across independent diploid extraction backgrounds. Such a signal is consistent with partial reproductive isolation or strong selection favoring particular allelic combinations when polyploid-derived chromosomes are returned to diploidy.

Marker distributions and ploidy classification

Across the joint diploid–tetraploid genotype calls, 10,102 markers were retained in S2039 and 5,945 in S2172. The smaller marker set in S2172 likely reflects lower effective polymorphism in that family, exacerbated by the smaller tetraploid sample.

Because variants were called under a tetraploid model for all individuals, we asked whether ploidy could be recovered solely from dosage profiles. Under ideal circumstances, diploid genotypes called as tetraploids should be confined to three dosage classes: nulliplex (0; homozygous reference), duplex (2; effective heterozygote), and tetraplex (4; homozygous alternative), with expected frequencies 0.25, 0.50, and 0.25, respectively. Simplex (1) and triplex (3) calls should be absent in true diploids. In the observed data, simplex and triplex calls are infrequent but not absent (3.1% and 2.7% in S2039D; 4.6% and 4.0% in S2172D), consistent with occasional misclassification of heterozygotes and providing a rough bound on genotype-calling error.

Nevertheless, the diploid populations show markedly lower proportions of simplex and triplex calls than their paired tetraploid populations (Figure 7). Motivated by this pattern, we summarized each individual by the proportions of markers called as simplex and triplex and used these two quantities as predictors in a linear discriminant analysis. Inspection of genotype-dosage histograms (Figure 7) indicated that simplex and triplex frequencies provide the clearest separation between ploidy classes.

The resulting discriminant function was

$$\text{Score} = -9.1856 + 41.6116 \times \text{prop}(\text{Simplex}) + 36.8495 \times \text{prop}(\text{Triplex}).$$

Discriminant scores ranged from -5.20 to -0.30 among diploids and from 4.04 to 12.07 among tetraploids, yielding complete separation of the two ploidy classes in this dataset. These results show that ploidy can be inferred with high accuracy from dosage-based summaries once GBS or similar SNP data are available, providing a practical complement to flow cytometry or chromosome counting in related interspecific sorghum populations.

Comparative genomic composition of diploid and tetraploid progeny

To further characterize relationships among individuals and ploidy classes, we computed modified Hamming distances from recoded genotypes and visualized the resulting distance matrices using multidimensional scaling. In both S2039 and S2172, diploid and tetraploid individuals separate cleanly along the first MDS axis (Figure 8). This separation reinforces the conclusion that the two ploidy classes are genetically distinct in each family and that sample mislabeling or seed contamination is unlikely to explain the observed structure.

Shared SNPs between diploid and tetraploid subsets enabled direct comparison of genomic composition via run-length analysis. In S2039D, five genomic intervals (on chromosomes 1, 3 [two regions], 5, and 6) exhibit runs of *S. bicolor*-fixed markers that exceed the significance threshold ($p < 0.05$, $k = 27$), while the same regions remain polymorphic in S2039T (Figure 3). These *S. bicolor*-fixed segments coincide with large map gaps, consistent with a lack of informative polymorphism in the diploids. The largest gap, on chromosome 1, spans 9.7 cM (approximately 4.6 Mb), and another on chromosome 6 spans 7.6 cM (about 4.7 Mb).

In S2172D, eight regions are fixed for *S. bicolor* alleles (chromosomes 1 [three regions], 3, 6, 7 [two regions], and 10; $p < 0.05$, $k = 13$), whereas the corresponding regions in S2172T remain segregating. Several of these segments align with large map gaps, including a region on chromosome 1 from 13.1–14.5 Mb that corresponds to a 37 cM gap, and a region on chromosome 4 from 53.7–56.6 Mb that coincides with a 15 cM gap. One fixed region on chromosome 3 lies within pericentromeric heterochromatin, suggesting that alternative alleles are retained in tetraploids but not in the derived diploids (Figure 5).

Across both families, we identified 13 genomic intervals that segregate in tetraploids but are fixed for *S. bicolor* alleles in diploids, a pattern unlikely to arise under neutral segregation given the overall allele frequencies ($p < 0.05$ genome-wide). These biases suggest that some *S. halepense*-derived alleles that are tolerated in a polyploid genomic context may be deleterious or strongly disfavored when returned to diploidy. Possible mechanisms include dosage sensitivity, subfunctionalization dependent on homoeologous gene copies, or epistatic interactions that require a polyploid genomic background.

Flowering time variation and QTLs across ploidy levels

Population means and standard deviations for flowering time are given in Table S1. In the S2039 families, tetraploids (S2039T) flower considerably later than diploids (S2039D), by 15.80 d in 2018 and 11.02 d in 2019, consistent with reports that polyploids often show delayed reproduction relative to diploids (e.g., [Petit et al. 1997](#)). In contrast, for S2172 the tetraploid population (S2172T) flowers earlier than the diploid S2172D by 6.36 d (2018) and 4.55 d (2019), illustrating that the direction of ploidy effects on flowering time is highly dependent on the allelic composition of the pedigree.

In S2039D, three QTLs for flowering time (qFL6.1, qFL9.1, and qFL10.1) exceed the permutation-based threshold, with qFL9.1 and qFL10.1 detected in both years (Table 4). In S2039T, single-marker scans identify two prominent association peaks on chromosome 6 (near 1.0 Mb and 47.1 Mb) but do not show strong signals near the diploid QTLs on chromosomes 9 or 10 (Figure 9). The S2039D QTL qFL6.1 maps roughly 9 Mb away from the dominant S2039T signal on chromosome 6, leaving open whether these signals reflect distinct loci or a shared causal region that is poorly resolved in one or both populations. Overall, the results suggest that flowering

time in S2039 is governed by different combinations of loci in diploid and tetraploid backgrounds.

The S2172D population harbors six flowering-time QTLs, five of which are detected in both years (Table 4). The largest-effect locus, qFL9.2, explains 26.09% of the phenotypic variance in 2018 and 10.53% in 2019, and maps approximately 7 Mb from qFL9.1 in S2039D. Interestingly, the chromosome 9 QTLs show opposite additive effects across families: in S2039D, the *S. halepense* allele is associated with earlier flowering, whereas in S2172D it is associated with delayed flowering. This sign reversal points to strong background dependence in the phenotypic expression of alleles in this region.

No flowering-time QTLs in S2172T meet the significance threshold, which is not surprising given the small tetraploid sample size ($n = 54$). Notably, however, loci corresponding to the major diploid QTLs qFL9.2 and qFL10.1 do not show conspicuous association peaks in the tetraploid data. Taken together with the S2039 results, this suggests that even when diploid and tetraploid families share parental genomes, the genetic control of flowering time can reconfigure substantially when ploidy changes.

Several of the flowering-time QTLs identified here co-localize with regions highlighted in the Sorghum QTL Atlas ([Mace et al. 2019](#)). In particular, QTLs on chromosomes 9 and 10 map near candidate genes including *ELF3*- and *CONSTANS*-like loci, which have been implicated in sorghum photoperiod sensitivity and floral initiation ([Li et al. 2018](#); [Yang et al. 2014](#)). These positional correspondences strengthen confidence in the QTL signals and highlight the utility of diploid extractions for connecting classical flowering-time loci to variation present in polyploid-derived germplasm.

Implications for diploidization, reproductive isolation, and breeding

Diploid families recently derived from polyploid progenitors are uncommon, and the S2039D and S2172D populations provide a rare window into transmission genetics at the onset of post-polyploid diploidization. The diploid progenitors of *S. halepense* are estimated to have diverged 1–2 million years ago ([Paterson et al. 1995](#)), placing the origin of the polyploid squarely within the timescale over which biased fractionation, structural change, and gene loss are expected following whole-genome duplication ([Lynch and Conery 2000](#); [Mandáková and Lysak 2018](#)). Comparisons between diploid and tetraploid populations derived from the same tetraploid parents suggest that some alleles captured in the polyploid are poorly tolerated or strongly selected against when returned to diploidy.

The shared enrichment of *S. halepense* alleles and segregation distortion on chromosome 6, together with multiple *S. bicolor*-fixed regions that segregate in tetraploids but not in diploids, points to a continuum of evolutionary outcomes. At one end are alleles that appear compatible across ploidy levels and introgress readily; at the other are alleles that may be dosage-sensitive, subfunctionalized, or embedded in polyploid-specific regulatory networks that break down in a diploid context. The mixed pattern of retention and loss we observe is consistent with a model in which polyploid *S. halepense* has accumulated novel combinations of alleles and regulatory interactions, some of which are effectively “locked in” to a polyploid genome organization.

These findings have both conceptual and applied consequences. From an ecological and biosafety perspective, *S. halepense* is often cited as a classic case of crop–weed gene flow and a potential route for transgene escape. Genomic features that reduce the fitness of

particular *S. bicolor* × *S. halepense* genotypes, or that restrict the viability of certain allele combinations in diploids, may mitigate such risks, whereas polyploid-derived alleles that confer strong fitness advantages (e.g., rhizomatousness or reproductive barriers) could be harnessed in strategic containment schemes (Gressel 2015).

At the same time, the diploid extractions studied here offer a practical avenue to exploit the evolutionary novelty of *S. halepense* for crop improvement. Traits such as perenniality, altered flowering behavior, yield components, and possibly interactions with soil biota can now be studied in a diploid framework that is directly compatible with elite sorghum breeding (Glover *et al.* 2010; Mace *et al.* 2019; Wang *et al.* 2015). The ability to distinguish ploidy reliably from SNP data, combined with clonable diploid lines that capture segments of the halepense genome, establishes a platform for systematically testing which polyploid-derived alleles are beneficial, neutral, or deleterious when introduced into modern sorghum backgrounds.

Data Availability

Raw genotyping-by-sequencing (GBS) FASTQ files, filtered SNP datasets for all four populations (S2039D, S2039T, S2172D, and S2172T), and flowering-time phenotype data from the 2018–2019 field seasons are deposited in the NCBI Sequence Read Archive (SRA) under BioProject accession **PRJNAxxxxxx**. All computational workflows used in this study—including genotype calling, SNP filtering, run-length analysis, ploidy-discriminant modeling, genetic map construction, and QTL mapping—are available at <https://github.com/<your-lab>/<repository-name>>.

Processed genetic maps, multidimensional scaling (MDS) distance matrices, and single-marker association statistics are provided as Supplementary Files accompanying this G3 submission. All datasets and scripts required to reproduce the analyses presented here are fully and freely accessible.

Acknowledgments

We thank the field and laboratory teams at The Land Institute and the University of Georgia for their essential contributions to planting, phenotyping, sample processing, and data collection across multiple seasons. We are also grateful to staff who supported greenhouse management, field logistics, and sample handling. Finally, we thank our colleagues and collaborators for valuable discussions that improved the design, analysis, and interpretation of this work.

Author Contributions

Conceptualization: S.C., A.H.P.

Methodology: W.K., P.N., S.C., J.R., A.H.P.

Formal Analysis: W.K., J.R.

Investigation: W.K., P.N., S.C., J.R., A.H.P.

Data Curation: W.K., P.N.

Visualization: W.K.

Writing – Original Draft: W.K., A.H.P.

Writing – Review & Editing: W.K., P.N., S.C., J.R., A.H.P.

Supervision: A.H.P., S.C.

Funding Acquisition: A.H.P., S.C.

Acknowledgment of Large Language Model Assistance

Portions of manuscript organization and language refinement were assisted through the use of ChatGPT (OpenAI, GPT-5.1 model,

accessed 2025). All AI-assisted content was reviewed, verified, and approved by the authors, who assume full responsibility for the accuracy and originality of the final manuscript.

Funding

This work was supported by a grant from the Malone Family Land Preservation Foundation and by the Perennial Agriculture Project, a joint initiative of The Land Institute and the Malone Family Land Preservation Foundation. Additional institutional support was provided by The Land Institute and the University of Georgia.

Conflicts of Interest

The authors declare no conflicts of interest. The funding organizations had no role in study design, data collection and analysis, interpretation of results, decision to publish, or preparation of the manuscript.

Literature cited

- Aljanabi SM, Forget L, Dookun A. 1999. An improved and rapid protocol for the isolation of polysaccharide- and polyphenol-free sugarcane dna. *Plant Molecular Biology Reporter*. 17:281.
- Arriola PE, Ellstrand NC. 1996. Crop-to-weed gene flow in the genus sorghum (poaceae): Spontaneous interspecific hybridization between johnsongrass, sorghum halepense, and crop sorghum, s. bicolor. *American Journal of Botany*. 83:1153–1159.
- Arriola PE, Ellstrand NC. 1997. Fitness of interspecific hybrids in the genus sorghum: Persistence of crop genes in wild populations. *Ecological Applications*. 7:512–518.
- Broman KW, Wu H, Sen S, Churchill GA. 2003. R/qtl: Qtl mapping in experimental crosses. *Bioinformatics*. 19:889–890.
- Cox S, Nabukalu P, Paterson A, Kong W, Nakasagga S. 2018a. Development of perennial grain sorghum. *Sustainability*. 10:172.
- Cox S, Nabukalu P, Paterson AH, Kong W, Auckland S, et al.. 2018b. High proportion of diploid hybrids produced by interspecific diploid × tetraploid sorghum hybridization. *Genetic Resources and Crop Evolution*. 65:387–390.
- Glover JD, Reganold JP, Bell LW, Borevitz J, Brummer EC, et al.. 2010. Increased food and ecosystem security via perennial grains. *Science*. 328:1638–1639.
- Gressel J. 2015. Dealing with transgene flow of crop protection traits from crops to their relatives. *Pest Management Science*. 71:658–667.
- Kong W. 2017. *Genetic dissection of plant architecture and life history traits salient to climate-resilient sustainable intensification of agriculture*. Ph.D. thesis. University of Georgia. Athens, GA.
- Kong W, Jin H, Franks CD, Kim C, Bandopadhyay R, et al.. 2013. Genetic analysis of recombinant inbred lines for sorghum bicolor × sorghum propinquum. *G3: Genes | Genomes | Genetics*. 3:101–108.
- Kong W, Kim C, Zhang D, Guo H, Tan X, et al.. 2018. Genotyping by sequencing of 393 sorghum bicolor btx623 × is3620c recombinant inbred lines improves sensitivity and resolution of qtl detection. *G3 (Bethesda)*. 8:2563–2572.
- Kong W, Nabukalu P, Cox T, Goff VH, Robertson JS, et al.. 2021. Quantitative trait mapping of plant architecture in two bc₁f₂ populations of sorghum bicolor × s. halepense and comparisons to two other sorghum populations. *Theoretical and Applied Genetics*. pp. 1–16.
- Kong W, Nabukalu P, Cox TS, Goff VH, Pierce GJ, et al.. 2020. Transmission genetics of a sorghum bicolor × s. halepense backcross population. *Frontiers in Plant Science*. 11:Article 1234.

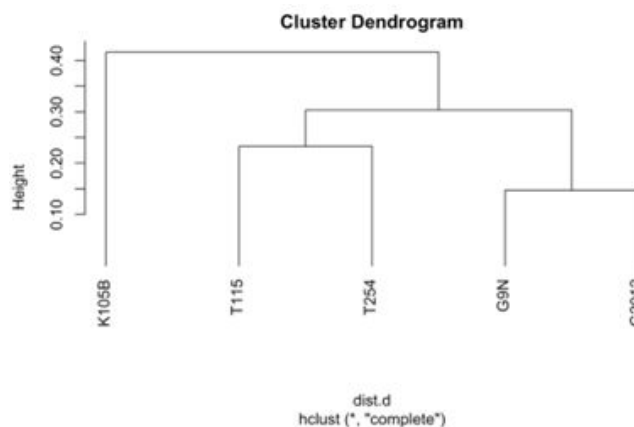


Figure 2 Hierarchical clustering of parental genotypes used to generate the S2039 and S2172 families. K105 is the cultivated *Sorghum bicolor* line; G9N and G2012 are two *S. halepense* accessions; T115 is the tetraploid parent of the S2172 populations; and T254 is the tetraploid parent of the S2039 populations.

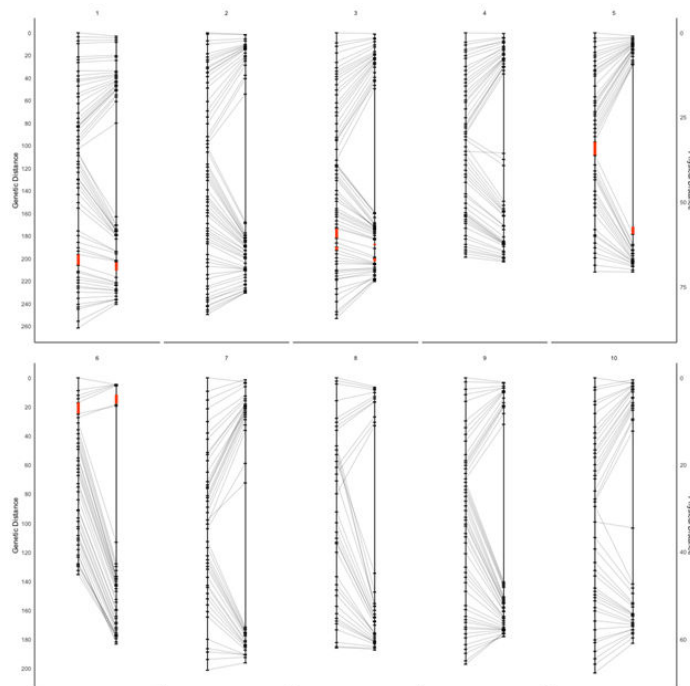


Figure 3 Genetic (left) and physical (right) maps for the S2039D population. Red segments indicate genomic regions showing non-random blocks of enrichment for *S. bicolor* alleles in S2039D relative to the corresponding tetraploid population S2039T.

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

Yang SS, Weers BD, Morishige DT, Mullet JE. 2014. Constans is a photoperiod-regulated activator of flowering in sorghum. *BMC Plant Biology*. 14:148.

- Lander ES, Botstein D. 1989. Mapping mendelian factors underlying quantitative traits using rflp linkage maps. *Genetics*. 121:185–199.
- Li H, Durbin R. 2009. Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics*. 25:1754–1760.
- Li X, Guo T, Mu Q, Li X, Yu J. 2018. Genomic and environmental determinants and their interplay underlying phenotypic plasticity. *Proceedings of the National Academy of Sciences*. 115:6679–6684.
- Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. *Science*. 290:1151–1155.
- Mace E, Innes D, Hunt C, Wang X, Tao Y, et al.. 2019. The sorghum qtl atlas: A powerful tool for trait dissection, comparative genomics, and crop improvement. *Theoretical and Applied Genetics*. 132:751–766.
- Mandáková T, Lysak MA. 2018. Post-polyploid diploidization and diversification through dysploid changes. *Current Opinion in Plant Biology*. 42:55–65.
- Margarido GR, Souza AP, Garcia AAF. 2007. Onemap: software for genetic mapping in outcrossing species. *Hereditas*. 144:78–79.
- McCormick RF, Truong SK, Sreedasyam A, Jenkins J, Shu S, et al.. 2018. The sorghum bicolor reference genome: improved assembly, gene annotations, a transcriptome atlas, and signatures of genome organization. *The Plant Journal*. 93:338–354.
- McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, et al.. 2010. The genome analysis toolkit: A mapreduce framework for analyzing next-generation dna sequencing data. *Genome Research*. 20:1297–1303.
- Morrell PL, Williams-Coplin D, Bowers JE, Chandler JM, Paterson AH. 2005. Crop-to-weed introgression has impacted allelic composition of johnsongrass populations with and without recent exposure to cultivated sorghum. *Molecular Ecology*. 14:2143–2154.
- Pandit MK, Tan H, Bisht M. 2006. Polyploidy in invasive plant species of singapore. *Botanical Journal of the Linnean Society*. 151:395–403.
- Parisod C, Holderegger R, Brochmann C. 2010. Evolutionary consequences of autopolyploidy. *New Phytologist*. 186:5–17.
- Paterson AH, Schertz KF, Lin YR, Liu SC, Chang YL. 1995. The weediness of wild plants: molecular analysis of genes influencing dispersal and persistence of johnsongrass, sorghum halepense (l.) pers. *Proceedings of the National Academy of Sciences*. 92:6127–6131.
- Petit C, Lesbros P, Ge X, Thompson JD. 1997. Variation in flowering phenology and selfing rate across a contact zone between diploid and tetraploid arrhenatherum elatius (poaceae). *Heredity*. 79:31–40.
- Piper JK, Kulakow PA. 1994. Seed yield and biomass allocation in sorghum bicolor and f₁ and backcross generations of s. bicolor × s. halepense hybrids. *Canadian Journal of Botany*. 72:468–474.
- Poland JA, Brown PJ, Sorrells ME, Jannink JL. 2012. Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. *PLoS ONE*. 7:e32253.
- Ritter KB, Jordan DR, Chapman SC, Godwin ID, Mace ES, et al.. 2008. Identification of qtl for sugar-related traits in a sweet × grain sorghum recombinant inbred population. *Molecular Breeding*. 22:367–384.
- Wang X, Mace E, Hunt C, Cruickshank A, Henzell R, et al.. 2014. Two distinct classes of qtl determine rust resistance in sorghum. *BMC Plant Biology*. 14:366.
- Wang X, Wang J, Jin D, Guo H, Lee TH, et al.. 2015. Genome align-

Table 1 Distribution of allele-enrichment groups in diploid (D) and tetraploid (T) S2039 and S2172 populations. Groups are defined by alternative-allele frequency p : *S. bicolor* enriched ($g = 0, p < 0.25$), equal contribution ($g = 1, 0.25 \leq p \leq 0.75$), and *S. halepense* enriched ($g = 2, p > 0.75$). Values are marker counts.

Group	S2039D	S2039T	S2172D	S2172T
<i>S. bicolor</i> enriched ($g = 0$)	6401	1690	2273	1103
Equal contribution ($g = 1$)	3114	7922	3020	4701
Alternative-allele enriched ($g = 2$)	587	490	652	141
Total markers	10102	10102	5945	5945

Table 2 Joint distribution of tetraploid parental genotypes (T254 and T115) and the composite *S. halepense* variable G_{combined} . Counts are followed in parentheses by genome-wide proportions.

(A) T254 vs. G_{combined}

	$G_{\text{combined}} = 0$	$G_{\text{combined}} = 1$	Total
T254 = 0	996 (0.072)	2288 (0.165)	3284 (0.237)
T254 = 1	2217 (0.160)	8331 (0.600)	10548 (0.760)
T254 = 2	8 (0.001)	88 (0.006)	96 (0.007)
Total	3221 (0.233)	10707 (0.771)	13928 (1.000)

(B) T115 vs. G_{combined}

	$G_{\text{combined}} = 0$	$G_{\text{combined}} = 1$	Total
T115 = 0	1191 (0.086)	4006 (0.288)	5197 (0.374)
T115 = 1	1853 (0.133)	5806 (0.417)	7659 (0.550)
T115 = 2	159 (0.011)	863 (0.062)	1022 (0.073)
Total	3203 (0.230)	10675 (0.767)	13878 (1.000)

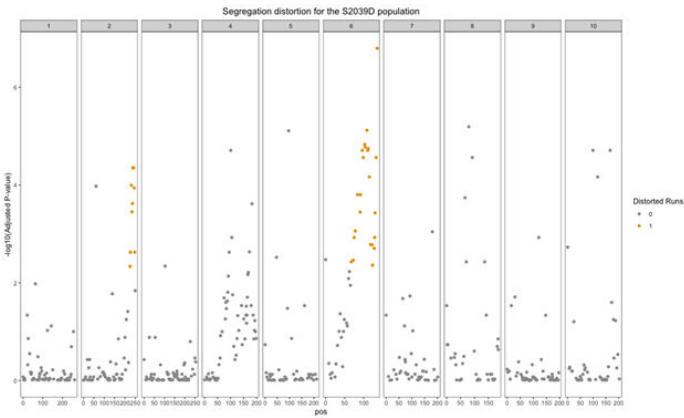


Figure 4 Segregation distortion patterns in the S2039D population. Yellow points highlight genomic positions with significant runs of distorted markers (more than 4 consecutive markers with $p < 0.05$), indicating local deviation from Mendelian segregation.

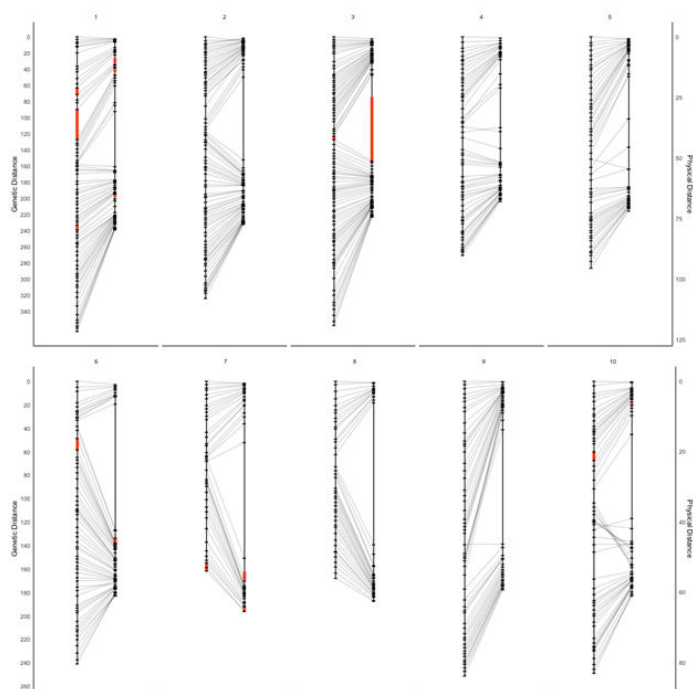


Figure 5 Genetic (left) and physical (right) maps for the S2172D population. Red segments indicate genomic regions showing non-random blocks of enrichment for *S. bicolor* alleles in S2172D relative to the corresponding tetraploid population S2172T.

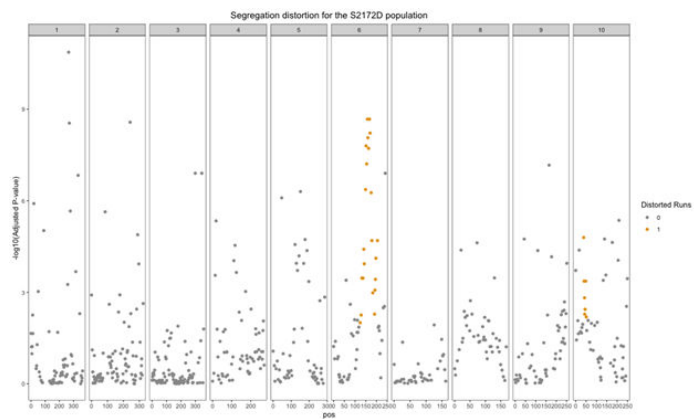


Figure 6 Segregation distortion patterns in the S2172D population. Yellow points highlight genomic positions with significant runs of distorted markers (more than 5 consecutive markers with $p < 0.05$), indicating local deviation from Mendelian segregation.

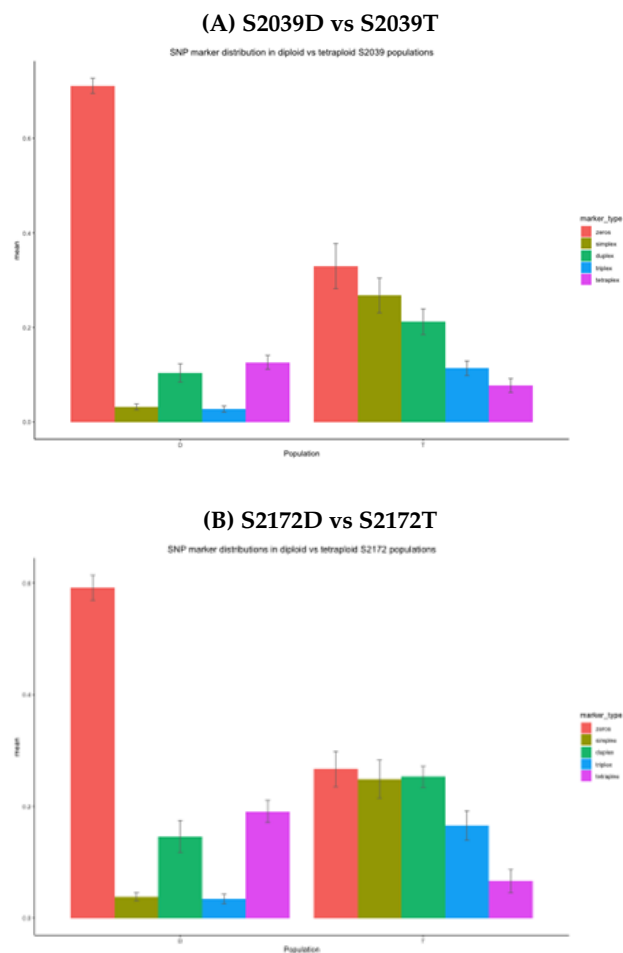


Figure 7 SNP dosage-class distributions in diploid versus tetraploid progeny. For each family, bars show the mean proportion of markers in each genotype class (nulliplex, simplex, duplex, triplex, tetraplex) under a tetraploid dosage-calling model, with error bars indicating standard errors across markers. Diploid populations show very low frequencies of simplex and triplex calls compared with their corresponding tetraploid populations, providing the basis for ploidy discrimination used in the linear discriminant analysis.

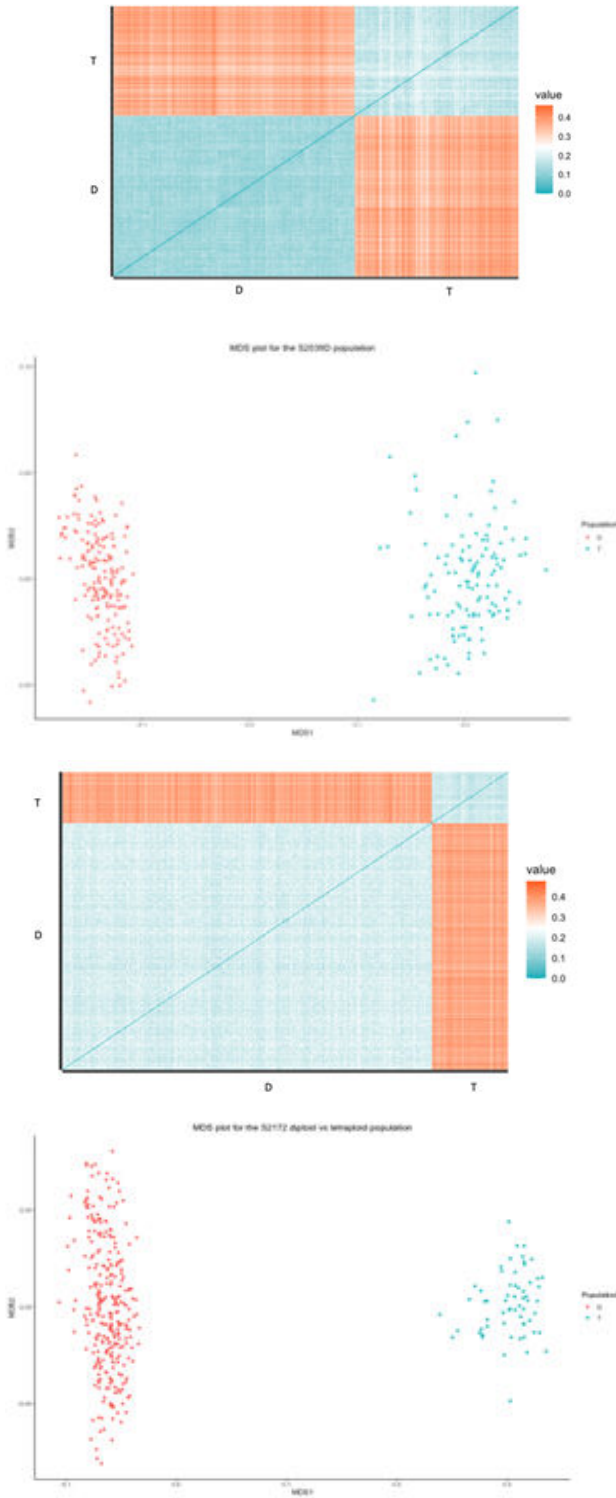


Figure 8 Distance matrices and multidimensional scaling (MDS) plots separating diploid and tetraploid progeny. (A) S2039D and S2039T. (B) S2172D and S2172T. For each family, the upper plot shows the pairwise modified Hamming distances between individuals, ordered so that diploid (D) and tetraploid (T) blocks are evident; warmer colors indicate larger distances. The lower plot shows the first two dimensions of an MDS analysis based on the same distance matrix. In both families, diploid and tetraploid individuals form clearly separated clusters, indicating strong genomic differentiation between ploidy classes and little evidence of seed contamination.

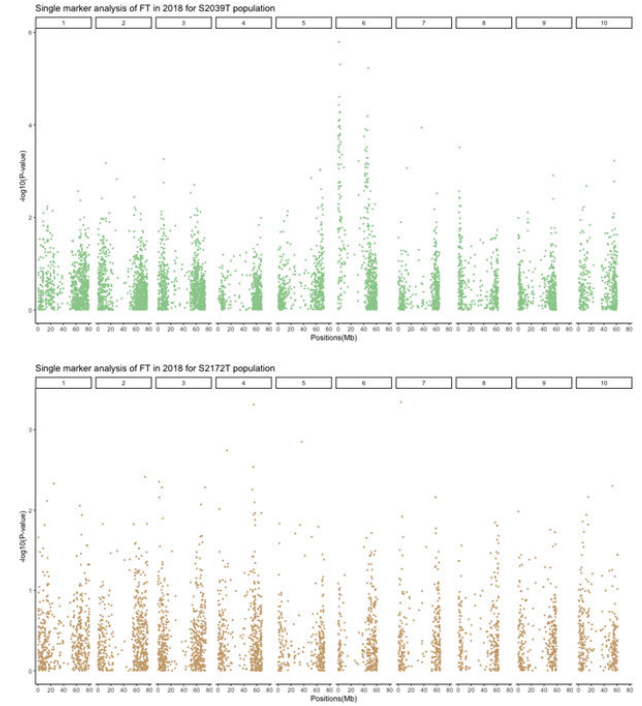


Figure 9 Single-marker association scans for flowering time (FT) in tetraploid populations using 2018 data. (A) S2039T. (B) S2172T. For each population, points show marker-wise $-\log_{10}(p)$ values plotted against physical position along the ten *Sorghum* chromosomes. Peaks on chromosome 6 in S2039T highlight candidate regions influencing flowering time, whereas no comparably strong signals are evident in S2172T, consistent with limited power due to smaller sample size.

Table 3 Summary of genetic map parameters for diploid S2039D and S2172D populations. Length and spacing are in centimorgans (cM). AA, AB, and BB denote genotype frequencies for homozygous reference, heterozygous, and homozygous alternative classes, respectively.

(A) S2039D (N = 180)

Chr	Markers (n)	Length (cM)	Avg spacing (cM)	Max spacing (cM)	AA (%)	AB (%)	BB (%)
1	60	261.2	4.4	20.4	27.16	47.61	25.22
2	63	249.5	4.0	11.9	30.44	47.52	22.04
3	64	252.8	4.0	9.5	25.16	49.58	25.26
4	55	198.7	3.7	10.5	32.34	47.73	19.93
5	48	211.4	4.5	12.5	26.87	47.71	25.42
6	42	135.4	3.3	8.4	16.79	48.50	34.71
7	43	200.9	4.8	15.7	28.82	45.78	25.40
8	36	185.6	5.3	16.7	25.82	48.46	25.72
9	51	197.0	3.9	13.4	28.02	49.50	22.48
10	42	202.9	4.9	10.9	27.90	46.45	25.65
Overall	504	2095.6	4.28	12.99	27.19 ^a	47.94	24.86

(B) S2172D (N = 300)

Chr	Markers (n)	Length (cM)	Avg spacing (cM)	Max spacing (cM)	AA (%)	AB (%)	BB (%)
1	73	364.2	5.1	37.0	27.30	45.80	26.89
2	84	323.5	3.9	15.8	28.11	46.20	25.69
3	90	356.7	4.0	15.0	26.36	48.16	25.48
4	59	270.3	4.7	15.3	30.98	46.65	22.37
5	53	285.9	5.5	12.9	29.45	47.86	22.69
6	61	240.8	4.0	10.0	22.29	46.78	30.93
7	38	161.8	4.4	11.2	25.51	51.10	23.39
8	42	167.7	4.1	10.4	22.25	45.23	32.52
9	54	251.1	4.7	11.8	28.90	46.38	24.72
10	56	249.0	4.5	22.6	30.46	44.97	24.57
Overall	610	2671.1	4.49	16.20	27.28	46.83	25.89

^aPercentages for the “Overall” row are weighted averages from raw genotype counts.

Table 4 Flowering time (FT) QTL detected in S2039D and S2172D populations across two years. LOD, location (Loc; cM), physical position (Pos; Mb), variance explained (Var; %), additive effect (*A*)^a, dominance effect (*D*)^b, and physical support interval (Start–End; Mb) are shown together with candidate genes or previously reported QTL and references.

Pop	Year	QTL	LOD	Loc (cM)	Pos (Mb)	Var (%)	<i>A</i>	<i>D</i>	Start (Mb)	End (Mb)	Candidate gene/QTL	Reference
S2039D	2018	qFL9.1	9.4	90.0	50.73	18.48	−1.74	−0.49	46.79	52.49	qFL9.1 (<i>S. propinquum</i>)	Kong <i>et al.</i> (2013)
S2039D	2018	qFL10.1	7.0	85.2	12.12	13.38	1.46	−0.26	7.41	47.25	CO	Yang <i>et al.</i> (2014)
S2039D	2019	qFL6.1	3.9	129.0	58.79	6.21	−1.06	0.36	57.76	59.34	qFL6.1 (IS3620C)	Kong <i>et al.</i> (2018)
S2039D	2019	qFL9.1	10.6	105.0	52.49	18.36	−1.82	−1.10	50.45	52.79	qFL9.1 (<i>S. propinquum</i>)	Kong <i>et al.</i> (2013)
S2039D	2019	qFL10.1	7.7	83.0	9.67	12.77	1.67	0.19	6.82	49.46	CO	Yang <i>et al.</i> (2014)
S2172D	2018	qFL1.1	7.5	160.3	55.05	6.60	−1.19	−1.11	53.49	55.83	qFL1.1 (IS3620C)	Kong <i>et al.</i> (2018)
S2172D	2018	qFL4.1	5.4	13.0	2.64	4.63	1.23	0.04	0.42	4.23	qFL4.1 (<i>S. propinquum</i>)	Kong <i>et al.</i> (2013)
S2172D	2018	qFL4.2	5.6	227.5	62.99	4.77	−1.29	0.92	62.42	63.23	qFL4D.H6.1 (<i>S. halepense</i>)	Kong <i>et al.</i> (2021)
S2172D	2018	qFL6.1	8.6	111.2	50.62	7.61	−1.67	0.13	50.18	50.94	–	–
S2172D	2018	qFL9.2	24.2	251.0	59.19	26.09	2.86	0.35	58.76	59.19	<i>Sobic.009G257300</i>	Li <i>et al.</i> (2018)
S2172D	2018	qFL10.1	12.1	68.0	6.85	11.25	1.92	−0.80	5.79	7.62	Maturity QTL	Ritter <i>et al.</i> (2008)
S2172D	2019	qFL4.1	4.3	39.3	3.32	5.98	1.33	0.05	2.68	5.95	qFL4.1 (<i>S. propinquum</i>)	Kong <i>et al.</i> (2013)
S2172D	2019	qFL4.2	4.0	197.5	61.96	5.56	−1.45	−0.22	59.50	64.34	qFL4D.H6.1 (<i>S. halepense</i>)	Kong <i>et al.</i> (2021)
S2172D	2019	qFL6.1	6.1	117.6	50.94	8.65	−1.66	−0.43	50.62	51.71	–	–
S2172D	2019	qFL9.2	7.4	249.0	59.19	10.53	1.84	0.12	57.81	59.19	<i>Sobic.009G257300</i>	Li <i>et al.</i> (2018)
S2172D	2019	qFL10.1	5.5	53.1	5.16	7.75	1.45	−1.40	4.61	8.53	QDTF_S7_10.1	Wang <i>et al.</i> (2014)

^aAdditive effect (*A*) is defined as half the difference between mean homozygotes.

^bDominance effect (*D*) is defined as the deviation of the heterozygote from the mid-point of homozygotes.