











Exceptional subgenome stability and functional divergence in the allotetraploid Ethiopian cereal teff

Robert VanBuren ^{1,2,9}✉, Ching Man Wai ^{1,2,9}, Xuewen Wang^{3,9}, Jeremy Pardo ^{1,2,4}, Alan E. Yocca ^{1,4}, Hao Wang³, Srinivasa R. Chaluvadi³, Guomin Han ³, Douglas Bryant⁵, Patrick P. Edger ¹, Joachim Messing ⁶, Mark E. Sorrells⁷, Todd C. Mockler ⁵, Jeffrey L. Bennetzen ³ & Todd P. Michael ⁸✉

Teff (*Eragrostis tef*) is a cornerstone of food security in the Horn of Africa, where it is prized for stress resilience, grain nutrition, and market value. Here, we report a chromosome-scale assembly of allotetraploid teff (variety Dabbi) and patterns of subgenome dynamics. The teff genome contains two complete sets of homoeologous chromosomes, with most genes maintaining as syntenic gene pairs. TE analysis allows us to estimate that the teff polyploidy event occurred ~1.1 million years ago (mya) and that the two subgenomes diverged ~5.0 mya. Despite this divergence, we detect no large-scale structural rearrangements, homoeologous exchanges, or biased gene loss, in contrast to many other allopolyploids. The two teff subgenomes have partitioned their ancestral functions based on divergent expression across a diverse expression atlas. Together, these genomic resources will be useful for accelerating breeding of this underutilized grain crop and for fundamental insights into polyploid genome evolution.

¹Department of Horticulture, Michigan State University, East Lansing, MI 48824, USA. ²Plant Resilience Institute, Michigan State University, East Lansing, MI 48824, USA. ³Department of Genetics, University of Georgia, Athens, GA 30602, USA. ⁴Department of Plant Biology, Michigan State University, East Lansing, MI 48824, USA. ⁵Donald Danforth Plant Science Center, St. Louis, MO 63132, USA. ⁶Waksman Institute of Microbiology, Rutgers University, Springfield, USA. ⁷Department of Plant Breeding and Genetics, Cornell University, Ithaca, NY, USA. ⁸J. Craig Venter Institute, La Jolla, CA 92037, USA. ⁹These authors contributed equally: Robert VanBuren, Ching Man Wai, Xuewen Wang. ✉email: bobvanburen@gmail.com; toddpmichael@gmail.com

Thirty crop species supply over 90% of the world's food needs and this narrow diversity reduces global food security. Humans have domesticated several hundred distinct plant species, but most are underutilized, under-improved, and restricted to their regions of origin¹. Although food systems have become increasingly diverse in the past few decades, many locally adapted species have been replaced by calorically dense staple crops, resulting in global homogeneity². Many underutilized and orphan crop species have desirable nutritional profiles, abiotic and biotic stress resilience, and untapped genetic potential for feeding our growing populations during this period of rapidly changing climate.

Teff is the staple grain crop in Ethiopia, and it is preferred over other cereals because of its nutritional profile, low input demand, adaptability, and cultural significance. Unlike other major cereals, teff is grown primarily by small-scale, subsistence farmers³ and thousands of locally adapted cultivars have been developed. Teff is among the most resilient cereals, tolerating marginal and semi-arid soils that are unsuitable for wheat, maize, sorghum, or rice production⁴. Teff was likely domesticated in the northern Ethiopian Highlands where much of the genetic diversity can be found^{5,6}. Consistent yields of small, nutritious seeds were the primary domestication targets of teff, contrasting most cereals where large seed heads and high productivity under tillage were desirable⁶. Despite its stress tolerance, yield improvements lag behind other cereals because of issues related to lodging, seed shattering, extreme drought, and poor agronomic practices⁷. Teff and other orphan cereals have undergone limited intensive selection for high productivity under ideal conditions, and rapid gains should be possible with advanced breeding and genome selection. A rough draft genome is available for the teff cultivar Tse dey (DZ-Cr-37)⁸, but the utility of these sequence data are limited because of the assembly's fragmented and incomplete nature.

The wild progenitor of teff is likely *Eragrostis pilosa*, a hardy wild grass sharing considerable overlap in morphological, genetic, and karyotype traits with teff^{9,10}. *Eragrostis tef* and *E. pilosa* are allotetraploids that arose from a shared polyploidy event that merged two currently unknown and possibly extinct or unsampled diploid genomes¹⁰. Many crop plants are polyploid, and genome doubling can give rise to emergent traits such as spinable fibers in cotton¹¹, morphological diversity in *Brassica* sp.¹², and new aromatic profiles of strawberry fruits¹³. Successful establishment of allopolyploids requires coordination of two distinct sets of homoeologous genes and networks, and often a dominant subgenome emerges to resolve genetic and epigenetic conflicts¹⁴. Newly formed polyploids are often unstable and undergo numerous structural rearrangements and fractionation compared to their diploid progenitors^{15,16}. Homeologous exchange is common during early polyploid formation, and large chromosome segments from one subgenome can replace another as observed in canola (*Brassica napus*), strawberry (*Fragaria ananassa*), cotton, and proso millet. The cotton allotetraploid complex formed around the same time as teff (1.7–1.9 million years ago (mya)), and the cotton A and D subgenomes diverged 6.2–7.1 mya. The two subgenomes have several hundred megabases of translocated sequences and structural rearrangements. This same pattern of rearrangement is observed in the banana (*Musa balbisiana*) A and B subgenomes, which diverged ~5.4 mya and have several megabase pair sized translocations and inversions between them. The allohexaploid false flax (*Camelina stativa*) has evidence of shattered chromosomes with numerous rearrangements and fractionation of the subgenomes compared to the diploid progenitors.

The effect of polyploidy on desirable traits and interactions between the two subgenomes remains untested in teff. Polyploidy

is found in more than 90% of species within the grass subfamily containing teff (Chloridoideae), and this has been hypothesized to contribute to the stress tolerance and diversification of these grasses¹⁷. Here, we report a chromosome-scale assembly of the teff A and B subgenomes and test for patterns of subgenome interactions and divergence.

Results

Genome assembly and annotation. We built a chromosome-scale assembly of the allotetraploid teff genome using a combination of long read single-molecule real-time sequencing and long-range high-throughput chromatin capture (Hi-C). In total, we generated 5.5 million filtered PacBio reads collectively spanning 52.9 Gb or 85× coverage of the estimated 622 Mb genome from the important teff landrace Dabbi. PacBio reads were error corrected and assembled using Canu¹⁸ and the resulting contigs were polished to remove residual errors with Pilon¹⁹ using high coverage Illumina data (45×). The PacBio assembly has a contig N50 of 1.55 Mb across 1344 contigs with a total assembly size of 576 Mb; 92.6% of the estimated genome size. The average nucleotide identity between homoeologous gene regions in teff is 93.9%, and this high sequence divergence facilitated accurate phasing and assembly. We utilized 20 random fosmids to assess the accuracy of the PacBio-based assembly (Supplementary Table 1). The fosmids collectively span 351 kb and have an average identity of 99.9% to the teff genome with individual fosmids ranging from 99.3 to 100%. This suggests that our assembly is mostly complete and accurately polished.

Contigs from the Canu-based draft genome were anchored into a chromosome-scale assembly using a Hi-C-based scaffolding approach. After filtering, 20 high-confidence clusters were identified, consistent with the haploid chromosome number of teff ($2n = 40$; Fig. 1). In total, 687 contigs collectively spanning 96% of the assembly (555 Mb) were anchored and oriented across the 20 pseudomolecules (Table 1). Pseudomolecules ranged in size from 19 to 40 Mb, consistent with the teff karyotype. Seven chimeric contigs corresponding to joined telomeres were

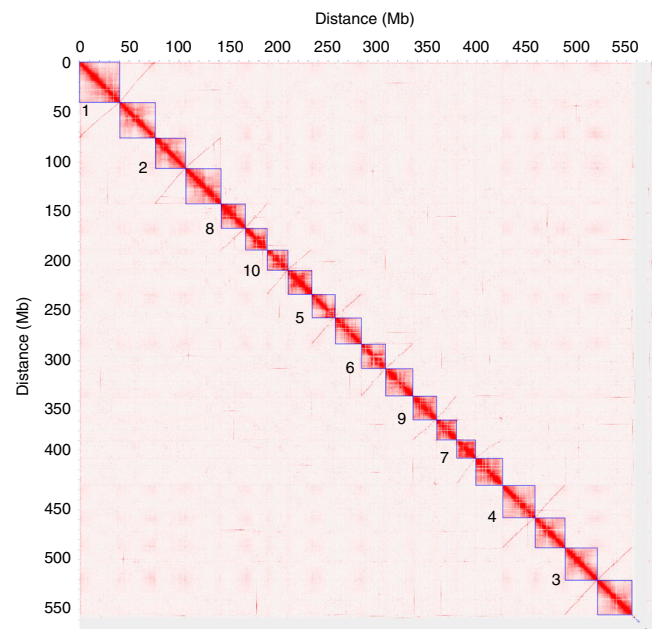


Fig. 1 Hi-C-based clustering of the teff genome. Heat map showing the density of Hi-C interactions between contigs, with red indicating high density of interactions. Distinct chromosomes are highlighted by blue boxes and homoeologous chromosome pairs are numbered.

Table 1 Summary statistics of the teff genome.

Chromosome	Size (bp)	Number of contigs	Number of genes	Number of tandem duplicates	Repetitive element content (%)
1A	40,621,098	35	5135	465	27.5
1B	35,710,944	32	4829	469	22.3
2A	35,425,885	45	4398	441	26.1
2B	30,633,641	23	4112	382	20.3
3A	34,643,735	47	4415	404	25.2
3B	32,575,812	43	4370	417	22.4
4A	32,664,196	39	4224	318	29.9
4B	29,936,223	32	4127	294	26.1
5A	26,945,638	29	2899	403	31.7
5B	24,206,550	36	2785	385	34.5
6A	27,140,163	46	2409	365	40.2
6B	19,415,607	31	1992	225	26.3
7A	26,459,500	44	3006	315	33.6
7B	23,383,462	34	2843	307	30.4
8A	24,151,120	26	2464	270	32.2
8B	21,147,804	28	2373	239	25.9
9A	24,589,398	38	2736	292	31.1
9B	21,940,566	23	2673	270	28.3
10A	23,813,772	24	2346	268	20.3
10B	20,101,091	32	2151	227	17.1
Unanchored	22,232,506	657	1968	130	18.2
Total	577,738,711	1344	68,255	6886	26.5

identified and split based on Hi-C interactions. This chromosome scale version is referred to as *teff* V3.

We assessed the accuracy of the pseudomolecule construction using a high-density single-nucleotide polymorphism (SNP)-based genetic map with 2002 markers across 32 linkage groups. This map was constructed using a recombinant inbred population derived from an interspecific cross of *E. tef* and *E. pilosa*²⁰. The pseudomolecules and genetic map are highly collinear with an average Pearson's correlation coefficient between marker and physical distance of $\rho = 0.932$ (Fig. 2). Several chromosomes are broken into multiple linkage groups because of low marker density and these linkage groups can be joined based on physical location on the pseudomolecules. There were some marker incongruences between homeologous linkage groups, but this was generally low, suggesting that the *teff* A and B subgenomes are accurately phased and assembled.

The *teff* genome was annotated using the MAKER pipeline. Transcript support from a large-scale expression atlas and protein homology to *Arabidopsis* and other grass genomes were used as evidence for ab initio gene prediction. After filtering transposon-derived sequences, ab initio gene prediction identified 68,255 gene models. We assessed the annotation quality using the Benchmarking Universal Single-Copy Ortholog (BUSCO) Embryophyta dataset. The annotation contains 98.1% of the 1440 core Embryophyta genes and the majority (1210) are found in duplicate in the A and B subgenomes.

The *teff* cultivar Tse dey (DZ-Cr-37) was previously sequenced using an Illumina-based approach, yielding a highly fragmented draft genome with 14,057 scaffolds and 50,006 gene models⁸. The fragmented nature of this assembly and incomplete annotation hinders downstream functional genomics, genetics, and marker-assisted breeding of *teff*. We compared the Tse dey assembly with our Dabbi reference to identify cultivar-specific genes and differences in assembly quality. Only 30,424 (60.8%) of the Tse dey gene models had similarity (>95% sequence identity) to gene models in our Dabbi reference, including 9866 homeologous gene pairs. Only 20,208 (29.6%) of our Dabbi gene models had homology to Tse dey gene models. The remaining gene models were unannotated or unassembled in the Tse dey assembly. Only one-third of the Tse dey genome is assembled into

scaffolds large enough to be classified as syntenic blocks to Dabbi, which is an unavoidable artifact of the poor assembly quality and low contiguity. Because of the fragmented nature of the Tse dey assembly, we were unable to identify lineage-specific genes. Hence, the genomic resources presented here represent a significant advance over previous efforts.

Subgenome characteristics. *Teff* is an allotetraploid with unknown diploid progenitors, but the polyploidy event is likely shared with other closely related *Eragrostis* species¹⁰. Because the diploid progenitors are unknown and possibly extinct, we utilized the putative centromeric array sequences to distinguish the homeologous chromosomes from the A and B subgenomes of *teff*. Putative centromeric (SatT) repeat arrays in *teff* range from 3.7 to 326 kb in size for each chromosome and individual arrays contain 22 to 824 copies (Supplementary Table 2). We identified two distinct SatT arrays in *teff* (hereon referred to as SatTA and SatTB). SatTA and SatTB are the same length (159 bp) but have different sequence composition (Supplementary Fig. 1b). This element was previously identified in Illumina short read data from the *teff* variety Enantite²¹. Alignment of the consensus SatT arrays identified several distinguishing polymorphisms and a maximum likelihood phylogenetic tree separated the SatT arrays into two well-supported clades (Supplementary Fig. 1a). Each clade contains one member from each of the ten homeologous chromosome pairs and this classification likely represents differences in SatT array composition between the diploid progenitor species. This approach allowed us to accurately distinguish homeologous chromosome pairs from the A and B subgenomes and verifies the allopolyploid origin of *teff*. This result is independently confirmed by an analysis that investigated historical transposable element (TE) activity (see below).

The *teff* subgenomes have 93.9% sequence similarity in the coding regions, suggesting that either the polyploidy event was relatively ancient or that the progenitor diploid species were highly divergent²². To estimate the divergence time of the A and B subgenomes, we calculated *K_s* (synonymous substitutions per synonymous site) between homeologous gene pairs. *Teff* homeologs have a single *K_s* peak with a median of 0.15 (Supplementary

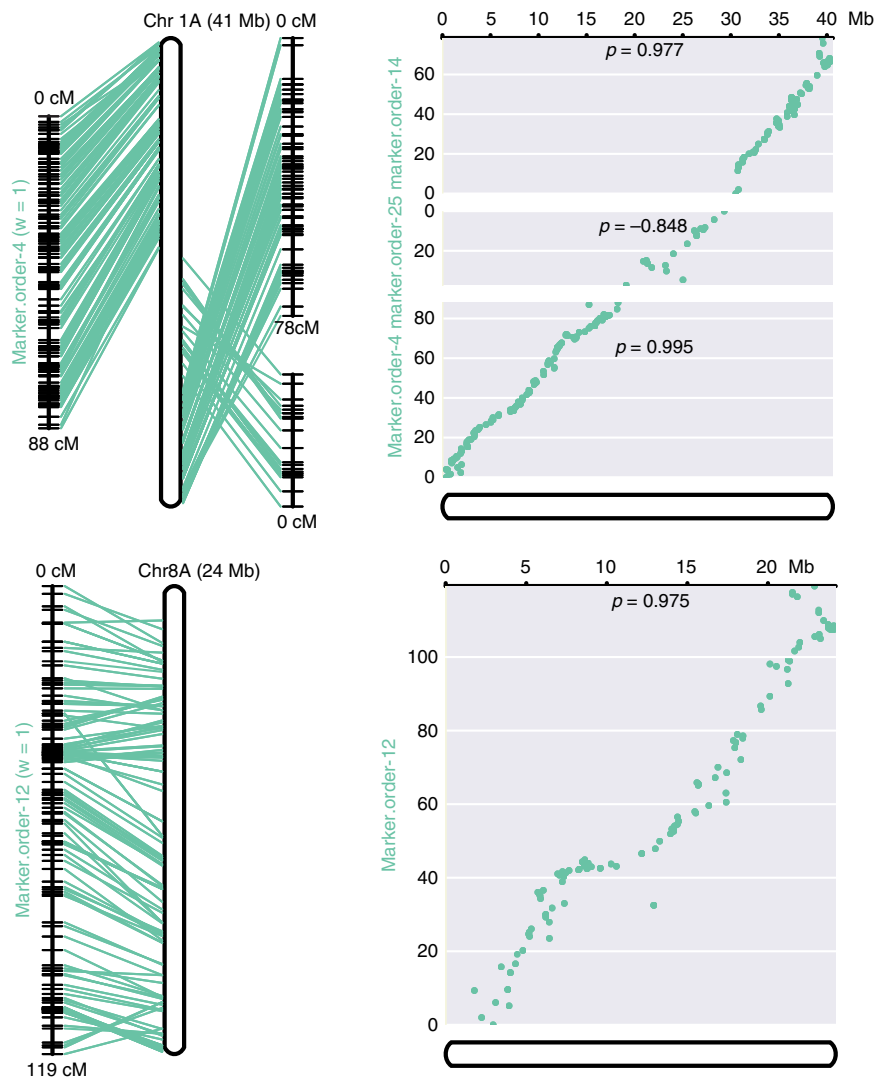


Fig. 2 Collinearity of *tef* pseudomolecules with the high-density genetic map. Two example chromosomes demonstrate a pseudomolecule spanning three linkage groups (top) and a pseudomolecule spanning a single linkage group (bottom). Lines connect the genetic makers with their physical location on the pseudomolecules. p Values within the scatterplots indicate the Pearson's correlation coefficient of marker distance (cM) and physical distance (bp). Source data are provided as a Source Data file.

Fig. 2), corresponding to a divergence time of ~ 5 million years based on a widely used mutation rate for grasses (1.5×10^{-8} substitutions per nonsynonymous site per year)²³. The ten pairs of homeologous chromosomes are highly syntenic with no large-scale structural rearrangements. The A subgenome is 13% (37 Mb) larger in size, but contains only 5% more genes than the B subgenome (34,032 vs. 32,255; Table 1). Most genes (54,846) are maintained as homeologous pairs and 13,409 are found in only one subgenome. Of these single-copy genes, 9036 have corresponding sequences in the homeologous chromosomes as either pseudogenes with frameshift mutations and missing exons, or low confidence gene models that were excluded from the final annotation. In total, $\sim 93.5\%$ of genes are maintained as homeologous gene pairs or a gene and pseudogene pair, with comparatively few being absent or deleted from one of the subgenomes. We identified 6876 tandemly duplicated genes with array sizes ranging from 2 to 15 copies. Of the 2748 tandem arrays, 998 are found in both subgenomes, while 864 and 1008 occur in only the A and B subgenomes, respectively (Table 1). Copy number varies extensively in shared arrays between the subgenomes.

The monoploid genome size of teff is relatively small (~ 300 Mb) compared to other polyploid grasses, and repetitive elements constitute a low percentage (25.6%) of the genome. Long terminal repeat-retrotransposons (LTR-RTs) are the most abundant repetitive elements, spanning at least 115.9 Mb or $\sim 20.0\%$ of the genome (Table 2). This predicted percentage is somewhat lower than that reported for other small grass genomes, such as *Oropetium* (250 Mb; 27%)²⁴ and *Brachypodium* (272 Mb; 21.4%)²⁵. We classified LTRs into 65 families and compared their abundance and insertion times (Fig. 3). A particular window of activity was seen for six families of LTR-RTs that were active only in the A genome progenitor or the B genome progenitor (Supplementary Fig. 3 and Supplementary Table 3). The insertion times for these genome-specific LTR-RTs were all greater than 1.1 mya, indicating the two subgenomes were evolving independently during this period. Hence, this LTR-RT analysis both confirms the A and B genome designations, and provides a methodology for determining the date of polyploid formation. In teff, these data indicate that the ancestral polyploidy was established ~ 1.1 mya.

Table 2 Summary of the repeat sequence distribution in the teff genome.

Class	Subclass	Superfamily	Number of families	Loci	Size (Mb)	Genome %
SSR	SSR	NA	1	116,936	5.2	0.9
Class I	LTR	Gypsy	944	54,384	71.8	12.4
	LTR	Unknown	946	55,889	32.5	5.6
	LTR	Copia	330	13,571	11.6	2
	LINE	L1	37	2784	1.6	0.3
	LINE	I	5	17	0	-0
Class II	SINE	Unknown	109	14,909	2.4	0.4
	TIR	Tc1	793	81,715	14.9	2.6
	TIR	CACTA	266	25,197	4.4	0.8
	TIR	hAT	77	7084	1.4	0.2
	TIR	PIF	48	5746	1.1	0.2
	TIR	Mutator	26	3238	0.6	0.1
	TIR	Unknown	1	247	0	-0
	Helitron	Helitron	105	21,977	5.6	1
	Total				153.1	26.5

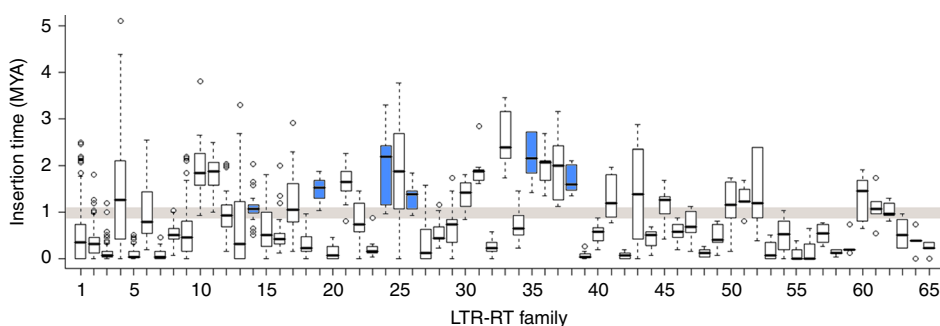


Fig. 3 Insertion dynamics of 65 LTR-RT families in teff. Box plots of insertion time for the 65 LTR-RT families having ≥ 5 intact LTR elements are plotted. Families 1–5 have ≥ 100 intact LTRs, 6–33 have ≥ 10 LTRs, and 34–65 have ≥ 5 LTRs. The exact number of LTR-RTs in each family is available in the TE annotation gff file. The six subgenome-specific families are highlighted in blue and the estimated range for the teff polyploidy event is shown in brown. A substitution rate of $1.3e-8$ per site per year was used to infer the element insertion times. Box boundaries indicate the 25th and 75th percentiles of the insertion time and whiskers extend to 1.5 times the interquartile range.

Five of the six subgenome-specific LTR-RT families were found only in the A subgenome, suggesting that LTR-RTs accumulate more rapidly in the A subgenome or are purged more effectively in the B subgenome. The A subgenome contains 35% more repetitive DNA than the B subgenome (87.5 vs. 64.9 Mb) and the recent bursts of LTR-RT activity contributes to the 13% larger size of the A subgenome. There are 24 families with median insertion times between 1.1 and 2.4 mya, and the remaining 18 families do not exhibit subgenomic specificity. Of these, 15 show no apparent burst in amplification, and three have evidence of very recent (post-polyploid) activity (Fig. 3, Supplementary Fig. 3, and Supplementary Table 4).

Teff belongs to the Chloridoideae subfamily of grasses, which includes important drought- and heat-tolerant C4 species such as the orphan grain crop finger millet and model desiccation tolerant plants in the genera *Oropetium*, *Eragrostis*, *Tripogon*, *Sporobolus*, and others. Most (~90%) of surveyed Chloridoideae species are polyploid, including many of the aforementioned taxa, and this likely contributes to their diversity and stress tolerance¹⁷. We utilized the wealth of genomic resources within Chloridoideae and more generally across Poaceae to identify patterns associated with improved stress tolerance, polyploidy, and genome evolution in teff. The teff and *Oropetium* genomes have a high degree of collinearity, as demonstrated by highly conserved gene content and order along each chromosome (Fig. 4). Teff and *Oropetium* show a clear 2:1 synteny pattern with 87% of teff genes having synteny to one block in *Oropetium* and 85% of *Oropetium* genes having synteny to two blocks in the teff genome (Fig. 4a). This

ratio corresponds to the A and B homoeologs of tetraploid teff and the single orthologs of diploid *Oropetium*. Each *Oropetium* chromosome has clear collinearity to two homoeologous teff chromosomes (Fig. 4c). Three trios have no rearrangements (teff 3A, 3B, and *Oropetium* Chr3; 4A, 4B, Chr4; 6A, 6B, Chr8) six trios have one or more large-scale inversions (1A, 1B, Chr1; 2A, 2B, Chr2; 5A, 5B, Chr7; 7A, 7B, Chr6; 8A, 8B, Chr9; 9A, 9B, Chr5) and one trio has translocations (10A, 10B, Chr10). Of the 28,909 *Oropetium* genes, 74% (21,293) have syntenic orthologs in both subgenomes of teff, 5% (1503) are found in only one subgenome, and 21% (6113) have no syntenic orthologs in teff. Teff and the allotetraploid grain crop finger millet have 2:2 synteny, but only 69% of syntenic blocks are found in duplicate because of the fragmented nature of the finger millet genome assembly²⁶ (Supplementary Fig. 4). Only 56% (38,149) of the teff genes have two syntenic orthologs in finger millet and the remaining 13 and 30% (9228 and 20,878) have one or zero syntenic orthologs in finger millet, respectively.

Following an allopolyploidy event, a dominant subgenome often emerges with significantly more retained genes and higher homoeolog expression as the plant returns to a diploid-like state²⁷. This dominance is established immediately following the polyploidy event, and patterns of biased fractionation have been observed in *Arabidopsis*²⁷, maize²⁸, *Brassica rapa*²⁹, and bread wheat³⁰. Biased homoeolog loss (fractionation) is not universal, and other allopolyploids such as *Capsella bursa-pastoris*³¹ and several *Cucurbita* species³² display no subgenome dominance. We searched for biased fractionation using syntenic orthologs

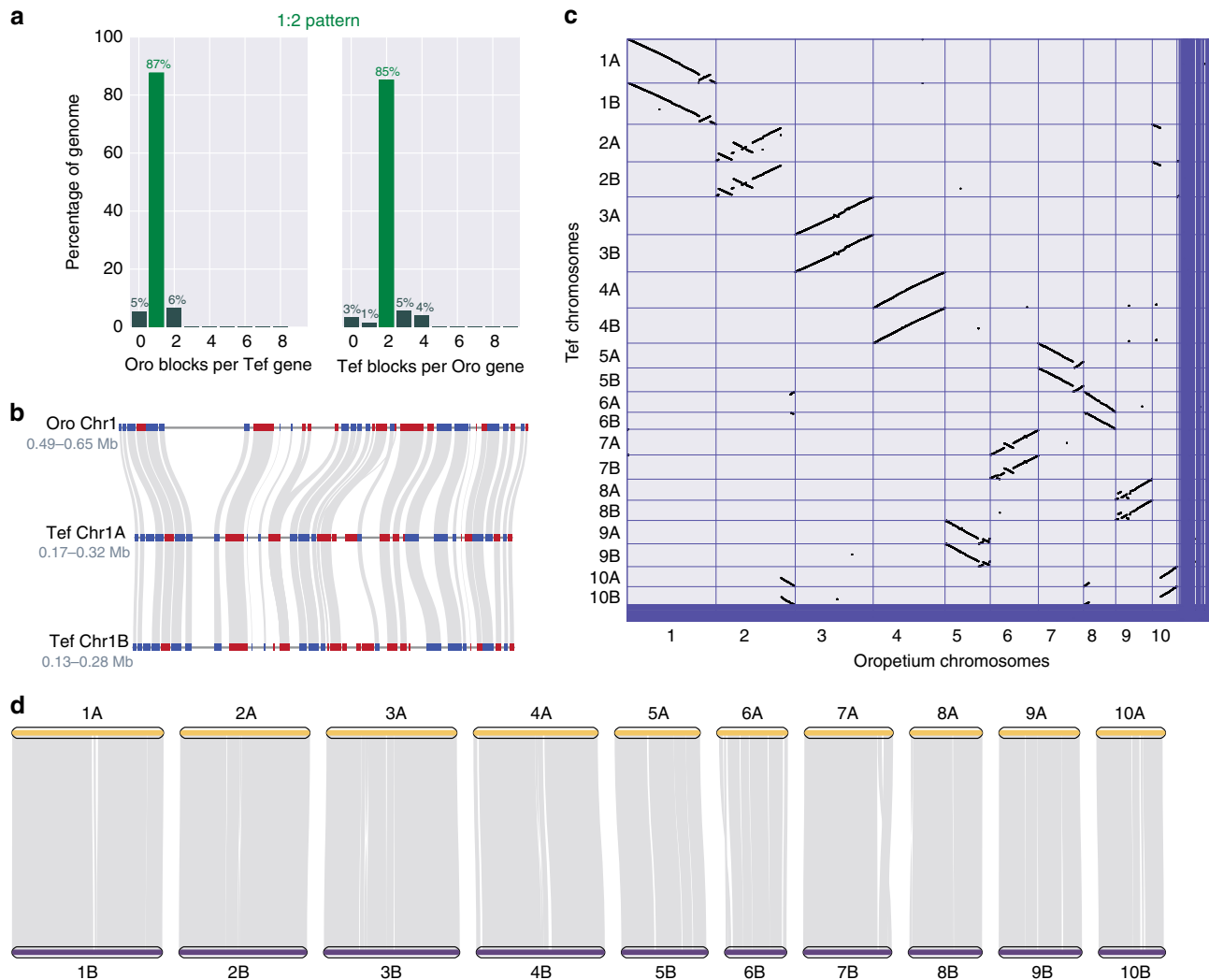


Fig. 4 Comparative genomics of the teff genome. **a** Ratio of syntenic depth between Oropetium and teff. Syntenic blocks of Oropetium per teff gene (left) and syntenic blocks of teff per Oropetium gene (right) are shown indicating a clear 1:2 pattern of Oropetium to teff. **b** Microsynteny of the teff and Oropetium genomes. A region of the Oropetium chromosome 1 and the corresponding syntenic regions in homoeologous teff chromosomes 1A and 1B are shown. Genes are shown in red and blue (for forward and reverse orientation, respectively) and syntenic gene pairs are connected by gray lines. **c** Macrosynteny of the teff and Oropetium genomes. Syntenic gene pairs are denoted by gray points. **d** Collinearity of the teff subgenomes. The ten chromosomes belonging to the teff A and B subgenomes are shown in yellow and purple, respectively. Syntenic blocks between homoeologous regions are shown in grey. Source data underlying Fig. 4c are provided as a Source Data file.

from Oropetium as anchors. The A and B subgenomes of teff have a near identical number of syntenic orthologs to Oropetium (21,697 vs. 21,520, respectively), suggesting that there is little or no biased fractionation in teff. Orthologs to 1325 Oropetium genes are found as single-copy loci in teff, including 647 and 678 from the A and B subgenomes, respectively. The remaining orthologs are maintained in duplicate in teff (21,276) compared to their single ortholog in Oropetium. Together, this suggests a general stability of gene content in *Eragrostis* after genome merger.

Homoeolog expression patterns and subgenome dominance.

To test for patterns of subgenome differentiation and dominance in teff, we surveyed gene expression in eight developmentally distinct tissue types and two stages of progressive drought stress. Sampled tissues include roots and shoots from seedlings and mature plants, internodes, and two stages of developing seeds. Tissue from mature, well-watered leaves and two time points of

severe drought were also collected (leaf relative water content of 33% and 16%, respectively). Of the 23,303 syntenic gene pairs between the A and B subgenomes, 15,325 have homoeologous expression bias (HEB) in at least one tissue, and 1694 have biased expression in all sampled tissues (Supplementary Fig. 5). Pairwise comparisons between syntenic gene pairs support a slight bias in transcript expression toward the B subgenome (Fig. 5a). Roughly 56% of the 207,873 pairwise comparisons across the ten tissues show biased expression toward homoeologs in the B subgenome. This pattern is consistently observed across all ten tissues and most chromosome pairs, but the difference is subtle when robust cutoffs of differential expression are applied (Fig. 5b, c; see Methods). Individual tissues have from 6061 to 8485 homoeologous gene pairs with significant differential expression, including 52.3% biased toward the B subgenome (Kruskal–Wallis H test $P < 0.01$; Fig. 5b). Eight pairs of chromosomes show HEB toward the B subgenome, and chromosomes 1 and 8 have more dominant homoeologs from the A subgenome, but the difference is not significant (Wilcoxon’s rank-sum $P > 0.05$). Together, this

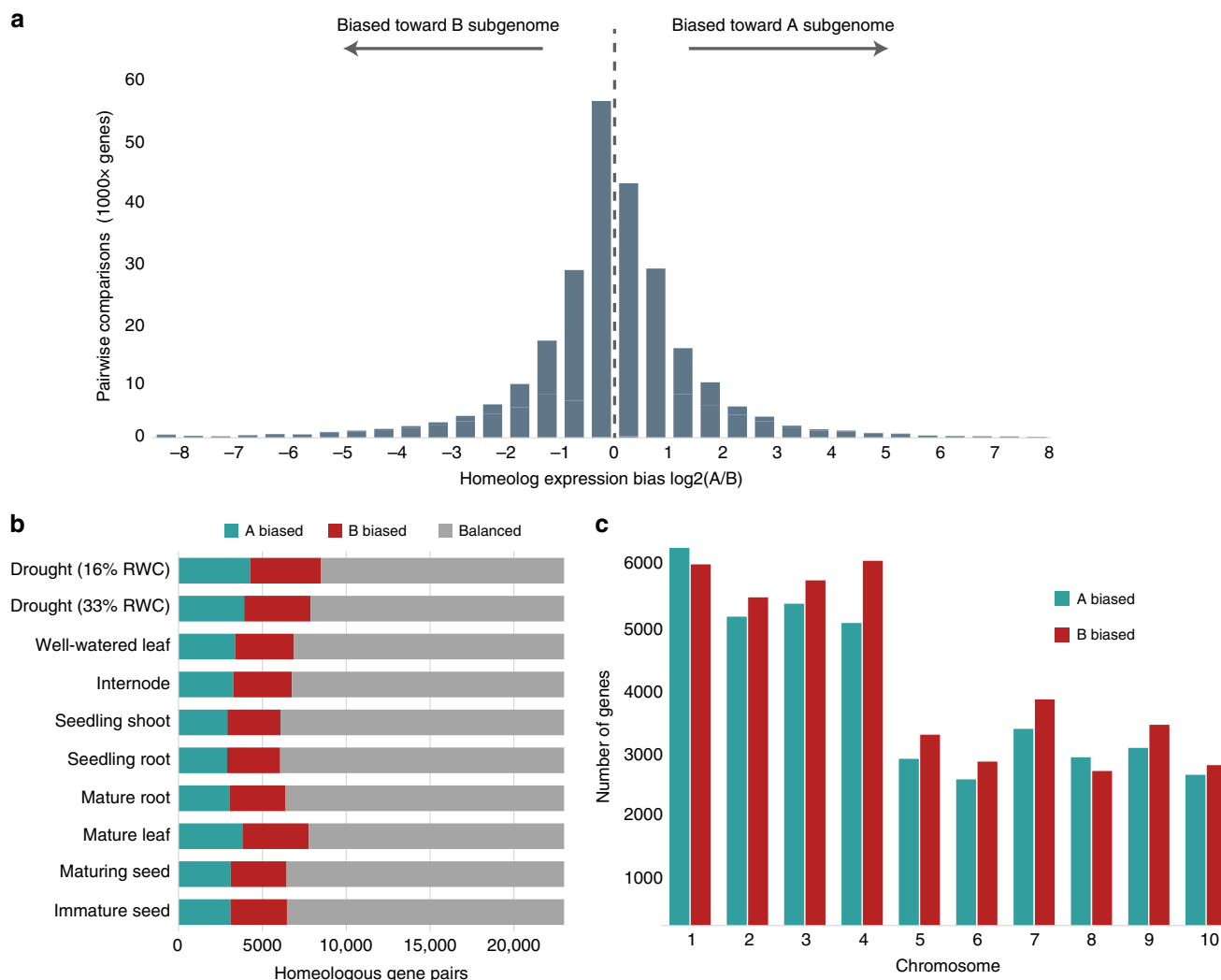


Fig. 5 Homoeolog expression bias between the A and B subgenomes of teff. **a** The distribution of homoeolog expression bias (HEB) between all gene pairs in all tissues. An HEB >0 indicates bias toward the A subgenome and a HEB <0 indicates bias toward the B subgenome. **b** HEB across the ten tissues in the teff expression atlas. Gene pairs were classified as biased toward the A (blue) or B (red) subgenomes or balanced with no statistically significant differential expression (gray). **c** HEB in each of the ten pairs of chromosomes across all ten tissue types. Source data underlying Fig. 5a are provided as a Source Data file.

suggests that the B subgenome is universally dominant over the A subgenome, but when strict thresholds are applied, this difference is minimal. Although we detected no evidence of recent homoeologous exchange, it is possible that genes from the recessive genome were replaced with homoeologs from the dominant subgenome during polyploid formation, which would weaken patterns of subgenome dominance.

One of the reasons for teff’s status as the preferred cereal in Ethiopia and among many food nutritionists is the high quality and quantity of proteins in the teff grain³³. Because teff seeds are so small, analyzing gene expression during seed development is challenging. For this reason, our investigations of the expression of teff storage protein genes during seed development included several early stages where we did not attempt to separate the developing seed from surrounding flower tissues (Supplementary Fig. 6). Hence, accurate quantification was not feasible, but analysis of the timing of expression was possible. The results indicated that the genes for all of the different classes of teff storage proteins, known as eragrostins³³, were primarily expressed during a narrow temporal window about 1 week after flower emergence (Supplementary Fig. 6). As for many self-pollinated cereals, we expect that the pollination occurred before

the flowers emerged from the stem at the boot leaf stage. For the most highly expressed storage protein genes (α -eragrostins on chromosomes 4A and 4B), there was a slight bias (53–47%) for expression from the A subgenome, while the next most highly expressed family for storage proteins (encoding the γ -eragrostins on chromosomes 2A and 2B), there was also a bias (82–18%) towards expression from the A subgenome.

We tested whether gene pairs with HEB maintain patterns of dominance across all tissues or whether dominant homoeologs are reversed in different tissues or under stress. The vast majority of genes (86.9%; 13,322) with HEB maintain the same pattern of dominance across all tissues, while 13.1% (2002) of the gene pairs have opposite dominance patterns in different tissues. The remaining 7675 gene pairs have no expression bias in any tissues or both homoeologs have negligible expression. Severely dehydrated leaf tissue had the most gene pairs with HEB (36%; 8485) compared to seedling roots and shoots, which each had ~26% of pairs with HEB. These results are consistent with previous findings in allohexaploid wheat³⁴ and allotetraploid *Tragopogon mirus*³⁵. We compared the ratio of nonsynonymous (Ka) to synonymous substitution rates (Ks) in homoeologous gene pairs to test if genes with stronger HEB are experiencing different

patterns of selection. Gene pairs with stronger HEB had significantly higher Ka/Ks than gene pairs with no HEB in any tissue (Supplementary Fig. 7; 0.17 vs. 0.28; Mann–Whitney $P < 0.01$). We detected no difference in divergence (Ks) among genes with varying degrees of HEB (Supplementary Fig. 8). Based on this pattern, we hypothesize that homoeologous gene pairs with higher expression divergence are under more relaxed selective constraints than gene pairs with balanced expression.

Discussion

Unlike the genomes of most polyploid grasses, the teff subgenomes are relatively small (~300 Mb), with high gene density and low TE content. The subgenomes are highly syntenic along their length, with no evidence of major inversions or structural rearrangements, in contrast to patterns observed in other relatively recent allopolyploids such as canola (*Brassica napus*)¹⁵, strawberry (*Fragaria ananassa*)¹⁶, and cotton³⁶. The cotton allotetraploid complex formed around the same time as teff (1.7–1.9 mya)³⁷, and the cotton A and D subgenomes diverged 6.2–7.1 mya. The two subgenomes have several hundred megabases of translocated sequences and structural rearrangements. This same pattern of rearrangement is observed in the banana (*Musa balbisiana*) A and B subgenomes, which diverged ~5.4 mya and have several megabase pair sized translocations and inversions between them³⁸. The allohexaploid false flax (*Camelina sativa*) has evidence of shattered chromosomes with numerous rearrangements and fractionation of the subgenomes compared to the diploid progenitors³⁹.

The general stability of the teff subgenomes may be attributed to low rates of homoeologous exchange. An estimated 90% of Chloridoideae grasses are polyploid and among the allopolyploid species, multivalent pairing is rarely detected¹⁷. The twenty chromosome pairs in teff show bivalent pairing in meiosis I, and double reduction has not been observed in segregating populations²⁰. Although homoeologous exchanges can result in advantageous emergent phenotypes, they can also destabilize the karyotype, leading to reduced fertility and fitness⁴⁰. For this reason, recent polyploids have long been considered evolutionary dead ends⁴¹. Thus, proper bivalent pairing (disomic inheritance) in natural allopolyploids may be favored, and the near perfect synteny observed between teff subgenomes suggests that an underlying mechanism may exist to prevent or reduce homoeologous exchanges in this species. We detected no evidence of recent homoeologous exchange in teff based on Ks distribution in windows across the genome, including exchanges that would have happened at the inception of the polyploidy event ~1.1 mya. Homoeologous exchanges are a common feature of allopolyploids, and the lack of these events is a unique feature of the teff genome.

The teff A and B subgenomes, and the Oropetium genome have high degrees of chromosome level collinearity despite their divergence. Oropetium and teff belong to different tribes within Chloridoideae (Cynodonteae and Eragrostideae, respectively) and diverged an estimated 25 mya^{42,43}. This is particularly unusual because polyploid-rich lineages typically have high rates of chromosome evolution⁴⁴. In contrast, our analysis of the divergence dates of the diploid A and B genome ancestors (~5 mya) and the formation of the tetraploid (~1.1 mya) indicates that the two genomes were so similar in structure (i.e., gene content, gene order and chromosome size) that some tetrasomic pairing would have been expected. Perhaps the *Ph1*-equivalent loci in *Eragrostis*, which control proper bivalent pairing in wheat⁴⁵, are so dominant that even low frequencies of homoeologous pairing are blocked. The high levels of subgenome compatibility, genetic and chromosome stability, fidelity for chromosome pairing, and low rates of homoeologous exchange allows polyploidy to dominate in the Chloridoideae subfamily. This polyploidy in turn may have

enabled the emergent resilience and robustness observed in Chloridoideae grasses.

Although we detected no biased fractionation between the teff subgenomes, we observed a general subgenome dominance across tissues in the expression atlas. The B subgenome is smaller and has fewer transposable elements, which may be contributing to the overall higher homoeolog expression levels. Patterns of B subgenome dominance are relatively weak compared to other allopolyploids, which may reflect the stability and lack of biased fractionation in teff. The teff subgenomes have successfully partitioned their ancestral roles, and most gene pairs display homoeolog expression bias. This bias is generally maintained across tissues and treatments, and few gene pairs change bias in a tissue-specific manner. Severely drought stressed leaf tissue has the highest proportion of genes with biased expression, which may reflect adaptation to adverse environments. Extensive homoeolog expression bias is also observed in hexaploid wheat³⁴ and tetraploid *Tragopogon mirus*³⁵ and may be a common feature of recent polyploid grasses.

The vast majority of genes in teff are maintained as homoeologous gene pairs in the A and B subgenomes, providing a significant obstacle for targeted breeding. Efforts to produce semi-dwarf, lodging-resistant teff using a mutagenesis approach have been more difficult because of gene redundancy⁴⁶. The resources provided here will help accelerate marker-assisted selection and guide genome engineering-based approaches, which must take gene redundancy into account. Most gene pairs have divergent expression profiles such that the subgenomes likely contribute unequally to different agronomic traits. Teff is often described as an orphan grain crop because of its limited investigation and improvement, resulting in relatively low yields under ideal conditions compared to other cereals with intensive selection and breeding histories. Teff and other grasses within Chloridoideae have high tolerance to abiotic stresses, and most of this resilience was maintained during teff domestication. This may represent a historical alternative selection scheme where maximum yield is exchanged for reliable harvest under poor environmental conditions. Future efforts to improve food security during rapidly changing climates should utilize the natural resilience of these robust, stable, polyploid species.

Methods

Plant materials. The Dabbi cultivar of teff (PI 524438, www.ars-grin.gov) was chosen for sequencing. Plant materials for high-molecular-weight (HMW) genomic DNA extraction, Hi-C library construction, and RNA were maintained in growth chambers under a 12-h photoperiod with day/night temperatures of 28 °C and 22 °C, respectively, and a light intensity of 400 $\mu\text{E m}^{-2} \text{s}^{-1}$. Tissue samples for the expression atlas were collected at ZT8 (Zeitgeber Time 8) to reduce issues associated with circadian oscillation. The Addisie cultivar of teff (PI 524434) was used for constructing the expression atlas. The tissue types used in the expression atlas include shoots and roots from young seedlings, mature leaf, internode, root, immature seeds, and mature seeds (Supplementary Fig. 9). For the drought time points, mature teff plants were allowed to dry slowly and leaf tissue was collected at subsequent days of extreme drought when the plant tissues had 33% and 16% relative water content, as well as well-watered teff for comparison. Tissue for the seed development timepoints were collected from teff florets at six different stages from 1 week before to 5 weeks after flower emergence. Three biological replicates were collected for each sample for RNAseq expression analysis. Leaf tissue from seedlings was used for the HMW genomic DNA extraction and Hi-C library construction. Tissues for HMW genomic DNA extraction and RNAseq were immediately frozen in liquid nitrogen and stored at –80 °C.

DNA library construction and sequencing. HMW genomic DNA was isolated from young teff leaf tissue for both PacBio and Illumina sequencing. A modified nuclei preparation⁴⁷ was used to extract HMW genomic DNA and residual contaminants were removed using phenol chloroform purification. PacBio libraries were constructed using the manufacturer's protocol and were size selected for 30 kb fragments on the BluePippen system (Sage Science), followed by subsequent purification using AMPure XP beads (Beckman Coulter). The PacBio libraries were sequenced on a PacBio RSII system with P6C4 chemistry. In total, 5.5 million filtered PacBio reads were generated, collectively spanning 52.9 Gb or ~85×

genome coverage (assuming a genome size of 622 Mb). The same batch of HMW genomic DNA was used to construct Illumina DNaseq libraries for correcting residual errors in the PacBio assembly. Libraries were constructed using the KAPA HyperPrep Kit (Kapa Biosystems), followed by sequencing on an Illumina HiSeq4000 under paired-end mode (150 bp).

RNA extraction and library construction. RNA for the expression atlas was extracted using the Omega Biotek E.Z.N.A.[®] Plant RNA Kit according to the manufacturer's protocol. Roughly 200 mg of ground tissue was used for each extraction. The RNA quality was validated using gel electrophoresis and the Qubit RNA IQ Assay (Thermo Fisher). Stranded RNAseq libraries were constructed using 2 µg of total RNA quantified using the Qubit RNA HS Assay Kit (Invitrogen, USA) with the Illumina TruSeq stranded total RNA LT Sample Prep Kit (RS-122-2401 and RS-122-2402). Multiplexed libraries were pooled and sequenced on an Illumina HiSeq4000 under paired-end 150 nt mode. Three replicates were sequenced for each timepoint/sample.

Genome assembly. The genome size of Dabbi teff was estimated using flow cytometry as previously described⁴⁸. The estimated flow cytometry size was 622 Mb, which was consistent with kmer-based estimations from Illumina data. The kmer plot had a unimodal distribution suggesting low within genome heterozygosity and high differentiation from the teff A and B subgenomes. Raw PacBio data was error corrected and assembled using Canu (v1.4)¹⁸, which produced accurate and contiguous assembly for homozygous plant genomes. The following parameters were modified: minReadLength = 2000, GenomeSize = 622 Mb, minOverlapLength = 1000. Assembly graphs were visualized after each iteration of Canu in Bandage⁴⁹ to assess complexities related to repetitive elements and homoeologous regions. The final Canu-based PacBio assembly has a contig N50 of 1.55 Mb across 1344 contigs with a total assembly size of 576 Mb. The raw PacBio contigs were polished to remove residual errors with Pilon (v1.22)¹⁹ using 73× coverage of Illumina paired-end 150 bp data. Illumina reads were quality-trimmed using Trimmomatic⁵⁰, followed by aligning to the assembly with bowtie2 (v2.3.0)⁵¹ under default parameters. Parameters for Pilon were modified as follows: --flank 7, --K 49, and --mindepth 15. Pilon was run recursively three times using the modified corrected assembly after each round. Ten full-length fosmids (collectively spanning 351 kb) were aligned to the final PacBio assembly to assess the quality. The fosmids exhibited an average identity of 99.9% to the PacBio assembly, with individual fosmids ranging from 99.3 to 100% nucleotide identity.

Genetic map construction. A previously generated recombinant inbred population derived from an interspecific cross of *E. tef* and *E. pilosa* was used to generate a high-density genetic map²⁰. GBS libraries were constructed with the ApeKI enzyme and sequenced on an Illumina HiSeq2000 sequencer under paired-end mode. The GBS reads were analyzed using teff contig assemblies as a reference, with the TASSEL-GBS pipeline (v4)⁵². Highly informative SNP markers (present in >80% of plants) were used for map construction. The genetic map was constructed using Joinmap (version 4.1)⁵³ and Mapmaker⁵⁴, using the Haldane function and a regression algorithm.

Hi-C analysis and pseudomolecule construction. The PacBio-based teff contigs were anchored into a chromosome-scale assembly using a Hi-C proximity-based assembly approach as previously described²⁴. A Hi-C library was constructed using 0.2 g of leaf tissue collected from newly emerged teff seedlings with the Proximo™ Hi-C Plant Kit (Phase Genomics) following the manufacturer's protocol. After verifying quality, the Hi-C library was size selected for 300–600 bp fragments and sequenced on the Illumina HiSeq4000 under paired-end 150 bp mode. One hundred and fifty base pair reads were used to avoid erroneous alignment in highly similar homoeologous regions. In total, 226 million read pairs were used as input for the Juicer and 3d-DNA Hi-C analysis and scaffolding pipelines^{55,56}. Illumina reads were quality-trimmed using Trimmomatic (0.39)⁵⁰ and aligned to the contigs using BWA (v0.7.16)⁵⁷ with strict parameters (-n 0) to prevent mismatches and non-specific alignments in repetitive and homoeologous regions. Contigs were ordered and oriented and assembly errors were identified using the 3d-DNA pipeline with default parameters⁵⁶. The resulting Hi-C contact matrix was visualized using Juicebox, and misassemblies and misjoins were manually corrected based on neighboring interactions. This approach identified 20 high-confidence clusters representing the haploid chromosome number in teff. The manually validated assembly was used to build pseudomolecules using the finalize-output.sh script from 3d-DNA and chromosomes were renamed and ordered by size and binned to the A and B subgenomes based on centromeric array analysis (described in detail below).

The accuracy of the Hi-C-based pseudomolecules was assessed using a high-density SNP-based genetic map with 2002 markers across 32 linkage groups. Several chromosomes were broken into multiple linkage groups in this map because of low population size. SNP-based markers were mapped to the teff genome using BLAST and collinearity between the physical and genetic map was assessed using the ALLMAPS package⁵⁸. The small differences between the pseudomolecules and genetic map are likely driven by missing data and marker distortion as well as the interspecific nature of this mapping population (*E. tef* × *E. pilosa*).

Identification of repetitive elements. We first identified and masked the simple sequence repeats in the teff genome with GMATA⁵⁹, and then conducted structure-based full-length TE identification using the following bioinformatic tools: LTR_retriever⁶⁰ to acquire high-confidence full LTR-RTs, SINE-Finder⁶¹ to identify SINEs, MGEscan-nonLTR (v2)⁶² to identify LINEs, MITE-Hunter⁶³ and MITE Tracker⁶⁴ to identify TIRs, and HelitronScanner⁶⁵ to identify *Helitrons*. All TEs were classified and manually checked according to the nomenclature system of transposons as described previously⁶⁶ and against Repbase to validate their annotation. We used the newly identified TEs as a custom library to identify full-length and truncated TE elements through a homology-based search with RepeatMasker (<http://www.repeatmasker.org>, version 4.0.7) using the teff pseudomolecules as input. The distribution of repeat sequences was then calculated. Only LTR-RT families with at least five intact copies were used for analysis of subgenome specificity. Within the 65 families having >5 intact elements, we identified LTRs with subgenomic specific activity. A family is considered as subgenomic specific if all intact elements of this family are from the same subgenome. Subgenome specificity was verified through BLAST of the element against the genome, and the distribution of matched sequences was manually inspected for subgenome specificity. The approximate insertion dates of LTR-RTs were calculated using the evolutionary distance between two LTR-RTs with the formula of $T = K/2\mu$, where K is the divergence rate approximated by percent identity and μ is the neutral mutation rate estimated as $\mu = 1.3 \times 10^{-8}$ mutations per bp per year²³.

Putative centromeric repeat arrays were identified with the approach outlined in ref. ⁶⁷ using Tandem repeat finder (version 4.07)⁶⁸. Parameters were modified as follows for tandem repeat finder: "1 1 2 80 5 200 2000 -d -h." Centromere-specific repeats are often the most abundant tandem repeats in the genome, and they were identified in teff by the following criteria: (1) copy number, (2) sequence level conservation between chromosomes, (3) similarity to other grass repeats, and (4) proximity to centromere-specific *gypsy* LTR-RTs. This approach identified two distinct centromere-specific arrays (SatTA and SatTB) with a shared length of 159 bp yet distinct sequence compositions. The consensus sequence of centromeric repeats from each chromosome was used to construct a maximum likelihood phylogenetic tree implemented in MEGA5 (v10.0.5)⁶⁹. This approach separated centromeric repeats from the 20 chromosomes into two distinct groups corresponding to the A and B subgenomes.

Genome annotation. Genes in the teff genome were annotated using the MAKER-P pipeline⁷⁰. The LTR-RT repeat library from LTR retriever was used for repeat masking. Transcript-based evidence was generated using RNAseq data from the ten tissues of the teff expression atlas. Quality-trimmed RNAseq reads were aligned to the unmasked teff genome using the splice aware alignment program STAR (v2.6)⁷¹ and transcripts were identified using StringTie (v1.3.4)⁷² with default parameters. The -merge flag was used to combine the output from individual libraries to generate a representative set of non-redundant transcripts. Protein sequences from the *Arabidopsis*, rice, and sorghum genomes as well as proteins from the UniProtKB plant databases⁷³ were used as protein evidence. Ab initio gene prediction was conducted using SNAP⁷⁴ and Augustus (3.0.2)⁷⁵ with two rounds of iterative training. The resulting gene models were filtered to remove any residual repetitive elements using BLAST with a non-redundant transposase library. The annotation quality was assessed using the BUSCO v2.76 with the plant-specific dataset (embryophyta_odb9).

Missing homoeologous genes were identified using the set of low confidence gene models that were purged from the final annotation because of insufficient evidence or characteristics of pseudogenes. BLAST was also used to identify any genes that were fragmented, pseudogenized, or were otherwise missed in the annotation. In total, evidence was found for 9036 homoeologous genes that were annotated as missing from one subgenome. Many of these genes had premature stop codons, were missing exons, or had no detectable expression, so most of these are presumably pseudogenes.

RNAseq expression analysis and homoeolog expression bias. Gene expression levels were quantified with the pseudo-aligner Kallisto (v0.44.0)⁷⁷ using the teff gene models as a reference. Paired-end Illumina reads from the ten tissues in the expression atlas were quality trimmed using Trimmomatic (v0.33) with default parameters and pseudo-aligned to the gene models with Kallisto under default parameters with 100 bootstraps per sample. The teff A and B subgenomes have high sequence divergence (~7%) such that misalignment between homoeologs was minimal. Expression levels were quantified as transcripts per million and the three biological replicates were averaged for direct homoeolog comparisons.

HEB was identified across all 1:1 homoeologous gene pairs using DESeq2⁷⁸ for all 10 samples in the teff expression atlas. Gene pairs were classified as having HEB if they passed the threshold of differential expression in DESeq2 using the following model (model 1):

$$y_{ij} \sim \mu + \text{timepoint} + e_{ij}. \quad (1)$$

Comparisons were made between the two homoeologous gene pairs for each of the ten samples in the atlas. The built-in Wald test in the DESeq2 package was used to test whether the log₂ fold change of a given gene was equal to 0 and genes with an false discovery rate-corrected, p value <0.05 were considered differentially expressed.

Comparative genomics. Homoeologous gene pairs between the teff A and B subgenomes and syntenic gene pairs across select grasses were identified using the MCSCAN toolkit implemented in python [[https://github.com/tanghaibao/jcvi/wiki/MCscan-\(Python-version\)](https://github.com/tanghaibao/jcvi/wiki/MCscan-(Python-version))]. Teff homoeologs were identified by all vs. all alignment using LAST, and hits were filtered using default parameters in MCSCAN with a minimum block size of five genes. Retained gene pairs from the ρ and σ whole-genome duplication events were filtered out using a C-score cutoff of 0.99. This approach identified 23,303 homoeologous, syntenic gene pairs between the A and B subgenome. Homoeologous gene pairs with translocations were not identified using this syntenic approach and were thus excluded from analysis. Tandem gene duplicates in teff were identified from the all vs. all LAST output with a maximum gene distance of 10. Gene models from teff were aligned to the *Oropetium thomaeum*²⁴ and *Sorghum bicolor* genes as outlined above for comparative genomics analyses across grasses. Macro and microsyntenic dot plots, block depths, and karyotype comparisons were generated in python using scripts from MSCAN.

Ka and Ks values were computed using a set of custom scripts available on GitHub: [https://github.com/Aeyocca/ka_ks_pipe/]. The homoeologous gene pair list from the teff subgenomes and syntenic orthologs between teff and *Oropetium* were used as input and the protein sequences from each gene pair were aligned using MUSCLE v3.8.31⁷⁹. PAL2NAL (v14)⁸⁰ was used to convert the peptide alignment to a nucleotide alignment and Ks values were computed between gene pairs using codeml from PAML (v4.9h) with parameters specified in the control file found in the GitHub repository listed above. Regions with recent homeologous exchanges were identified by comparing Ks values of syntenic gene pairs along windows across the genome.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

A reporting summary for this article is available as a Supplementary Information file. Data supporting the findings of this work are available within the paper and its Supplementary Information files. The datasets generated and analyzed during the current study are available from the corresponding author upon request. The raw PacBio data, Illumina DNaseq, and RNAseq data are available from the National Center for Biotechnology Information Short Read Archive under bioproject [PRJNA552060](https://www.ncbi.nlm.nih.gov/bioproject/PRJNA552060). RNAseq reads from the teff expression atlas were deposited to the National Center for Biotechnology Information Short Read Archive under bioproject [PRJNA525065](https://www.ncbi.nlm.nih.gov/bioproject/PRJNA525065). The genome assembly and annotation for teff is available from CoGe under [id50954](https://www.cbcb.umd.edu/coge/entry.do?entryId=50954). The UniProtKB plant databases [<https://www.uniprot.org/help/plants>] and embryophyta_odb9 BUSCO dataset [https://busco-archive.ezlab.org/v2/datasets/embryophyta_odb9.tar.gz] were downloaded from source for data analyses. The source data underlying Figs. 2, 4c, and 5a and Supplementary Figs. 2, 3, 4a, 5, 7, and 8 are provided as a Source Data file.

Code availability

Custom scripts for calculating Ka/Ks values of homeologous gene pairs are available on GitHub: [https://github.com/Aeyocca/ka_ks_pipe/].

Received: 28 June 2019; Accepted: 30 January 2020;

Published online: 14 February 2020

References

- Mueller, N. G., Fritz, G. J., Patton, P., Carmody, S. & Horton, E. T. Growing the lost crops of eastern North America's original agricultural system. *Nature Plants* **3**, 17092 (2017).
- Khoury, C. K. et al. Increasing homogeneity in global food supplies and the implications for food security. *Proc. Natl Acad. Sci. USA* **111**, 4001–4006 (2014).
- CSA. *Agricultural Sample Survey 2011/2012: Report on Area and Production of Major Crops* (Central Statistical Agency Addis Ababa, 2012).
- Stallknecht, G. F., Gilbertson, K. M. & Eckhoff, J. in *New crops* 231–234 (Wiley, New York, 1993).
- Demissie, A. in *Narrowing the Rift. Tef Research and Development. Proc. International Workshop on Tef Genetics and Improvement* (Debre Zeit, Ethiopia, 2000).
- D'Andrea, A. C. T'ef (*Eragrostis tef*) in ancient agricultural systems of highland Ethiopia. *Econ. Bot.* **62**, 547–566 (2008).
- Abraham, B. et al. The system of crop intensification: reports from the field on improving agricultural production, food security, and resilience to climate change for multiple crops. *Agric. Food Security* **3**, 4 (2014).
- Cannarozzi, G. et al. Genome and transcriptome sequencing identifies breeding targets in the orphan crop tef (*Eragrostis tef*). *BMC Genomics* **15**, 581 (2014).
- Gugs, L. et al. in *Narrowing the Rift: Tef Research and Development. Proc. International Workshop on Tef Genetics and Improvement* (Debre Zeit, Ethiopia, 2001).
- Ingram, A. L. & Doyle, J. J. The origin and evolution of *Eragrostis tef* (Poaceae) and related polyploids: evidence from nuclear waxy and plastid rps16. *Am. J. Bot.* **90**, 116–122 (2003).
- Paterson, A. H. et al. Repeated polyploidization of *Gossypium* genomes and the evolution of spinnable cotton fibres. *Nature* **492**, 423 (2012).
- Osborn, T. C. The contribution of polyploidy to variation in *Brassica* species. *Physiol. Plant.* **121**, 531–536 (2004).
- Ulrich, D. & Olbricht, K. Diversity of volatile patterns in sixteen *Fragaria vesca* L. accessions in comparison to cultivars of *Fragaria x ananassa*. *J. Appl. Bot. Food Qual.* **86**, 36–46 (2013).
- Freeling, M. et al. Fractionation mutagenesis and similar consequences of mechanisms removing dispensable or less-expressed DNA in plants. *Curr. Opin. Plant Biol.* **15**, 131–139 (2012).
- Chalhoub, B. et al. Early allopolyploid evolution in the post-Neolithic *Brassica napus* oilseed genome. *Science* **345**, 950–953 (2014).
- Edger, P. P. et al. Origin and evolution of the octoploid strawberry genome. *Nat. Genet.* **51**, 541–547 (2019).
- Roodt, R. & Spies, J. J. Chromosome studies in the grass subfamily Chloridoideae. II. An analysis of polyploidy. *Taxon* **52**, 736–746 (2003).
- Koren, S. et al. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).
- Walker, B. J. et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE* **9**, e112963 (2014).
- Yu, J.-K. et al. A genetic linkage map for tef [*Eragrostis tef* (Zucc.) Trotter]. *Theor. Appl. Genet.* **113**, 1093–1102 (2006).
- Gebre, Y. G., Bertolini, E., Pè, M. E. & Zuccolo, A. Identification and characterization of abundant repetitive sequences in *Eragrostis tef* cv. Enatite genome. *BMC Plant Biol.* **16**, 39 (2016).
- Doyle, J. J. & Egan, A. N. Dating the origins of polyploidy events. *N. Phytologist* **186**, 73–85 (2010).
- SanMiguel, P., Gaut, B. S., Tikhonov, A., Nakajima, Y. & Bennetzen, J. L. The paleontology of intergene retrotransposons of maize. *Nat. Genet.* **20**, 43–45 (1998).
- VanBuren, R., Wai, C. M., Keilwagen, J. & Pardo, J. A chromosome-scale assembly of the model desiccation tolerant grass *Oropetium thomaeum*. *Plant Direct* **2**, e00096 (2018).
- Initiative, I. B. Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature* **463**, 763–768 (2010).
- Hittalmani, S. et al. Genome and transcriptome sequence of finger millet (*Eleusine coracana* (L.) Gaertn.) provides insights into drought tolerance and nutraceutical properties. *BMC Genomics* **18**, 465 (2017).
- Thomas, B. C., Pedersen, B. & Freeling, M. Following tetraploidy in an *Arabidopsis* ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. *Genome Res.* **16**, 934–946 (2006).
- Schnable, J. C., Springer, N. M. & Freeling, M. Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *Proc. Natl Acad. Sci. USA* **108**, 4069–4074 (2011).
- Wang, X. et al. The genome of the mesopolyploid crop species *Brassica rapa*. *Nat. Genet.* **43**, 1035–1039 (2011).
- Li, A. et al. mRNA and small RNA transcriptomes reveal insights into dynamic homoeolog regulation of allopolyploid heterosis in nascent hexaploid wheat. *Plant Cell* **26**, 1878–1900 (2014).
- Douglas, G. M. et al. Hybrid origins and the earliest stages of diploidization in the highly successful recent polyploid *Capsella bursa-pastoris*. *Proc. Natl Acad. Sci. USA* **112**, 2806–2811 (2015).
- Sun, H. et al. Karyotype stability and unbiased fractionation in the paleo-allotetraploid *Cucurbita* genomes. *Mol. Plant* **10**, 1293–1306 (2017).
- Zhang, W., Xu, J., Bennetzen, J. L. & Messing, J. Teff, an orphan cereal in the chloridoideae, provides insights into the evolution of storage proteins in grasses. *Genome Biol. Evolution* **8**, 1712–1721 (2016).
- Ramirez-González, R. et al. The transcriptional landscape of polyploid wheat. *Science* **361**, eaar6089 (2018).
- Buggs, R. J. et al. Tissue-specific silencing of homoeologs in natural populations of the recent allopolyploid *Tragopogon mirus*. *N. Phytologist* **186**, 175–183 (2010).
- Wang, M. et al. Reference genome sequences of two cultivated allotetraploid cottons, *Gossypium hirsutum* and *Gossypium barbadense*. *Nat. Genet.* **51**, 224–229 (2018).
- Hu, Y. et al. *Gossypium barbadense* and *Gossypium hirsutum* genomes provide insights into the origin and evolution of allotetraploid cotton. *Nat. Genet.* **51**, 739–748 (2019).
- Wang, Z. et al. *Musa balbisiana* genome reveals subgenome evolution and functional divergence. *Nat. Plants* **5**, 810–821 (2019).

39. Mandáková, T., Pouch, M., Brock, J. R., Al-Shehbaz, I. A. & Lysak, M. A. Origin and evolution of diploid and allopolyploid *Camelina* genomes was accompanied by chromosome shattering. *Plant Cell* **31**, 2596–2612 (2019).
40. Gaeta, R. T. & Pires, J. C. Homoeologous recombination in allopolyploids: the polyploid ratchet. *N. Phytologist* **186**, 18–28 (2010).
41. Mayrose, I. et al. Recently formed polyploid plants diversify at lower rates. *Science* **333**, 1257–1257 (2011).
42. Christin, P.-A. et al. Oligocene CO₂ decline promoted C4 photosynthesis in grasses. *Curr. Biol.* **18**, 37–43 (2008).
43. Vicentini, A., Barber, J. C., Alicioni, S. S., Giussani, L. M. & Kellogg, E. A. The age of the grasses and clusters of origins of C4 photosynthesis. *Glob. Change Biol.* **14**, 2963–2977 (2008).
44. Wendel, J. F. in *Plant Molecular Evolution* (Springer, 2000).
45. Riley, R. & Chapman, V. Genetic control of the cytologically diploid behaviour of hexaploid wheat. *Nature* **182**, 713–715 (1958).
46. Zhu, Q. et al. High throughput discovery of mutations in *tef* semi-dwarfing genes by next generation sequencing analysis. *Genetics* **192**, 819–829 (2012).
47. Zhang, H. B., Zhao, X., Ding, X., Paterson, A. H. & Wing, R. A. Preparation of megabase-size DNA from plant nuclei. *Plant J.* **7**, 175–184 (1995).
48. Arumuganathan, K. & Earle, E. Estimation of nuclear DNA content of plants by flow cytometry. *Plant Mol. Biol. Rep.* **9**, 229–241 (1991).
49. Wick, R. R., Schultz, M. B., Zobel, J. & Holt, K. E. Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics* **31**, 3350–3352 (2015).
50. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
51. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
52. Glaubitz, J. C. et al. TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. *PLoS ONE* **9**, e03046 (2014).
53. Stam, P. Construction of integrated genetic linkage maps by means of a new computer package: Join Map. *Plant J.* **3**, 739–744 (1993).
54. Lander, E. S. et al. MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. *Genomics* **1**, 174–181 (1987).
55. Durand, N. C. et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* **3**, 95–98 (2016).
56. Dudchenko, O. et al. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
57. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at <http://arxiv.org/abs/1303.3997> (2013).
58. Tang, H. et al. ALLMAPS: robust scaffold ordering based on multiple maps. *Genome Biol.* **16**, 3 (2015).
59. Wang, X. & Wang, L. GMATA: an integrated software package for genome-scale SSR mining, marker development and viewing. *Front. Plant Sci.* **7**, 1350 (2016).
60. Ou, S. & Jiang, N. LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol.* **176**, 1410–1422 (2018).
61. Wenke, T. et al. Targeted identification of short interspersed nuclear element families shows their widespread existence and extreme heterogeneity in plant genomes. *Plant Cell* **23**, 3117–3128 (2011).
62. Rho, M. & Tang, H. MGEScan-non-LTR: computational identification and classification of autonomous non-LTR retrotransposons in eukaryotic genomes. *Nucleic Acids Res.* **37**, e143 (2009).
63. Han, Y. & Wessler, S. R. MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Res.* **38**, e199 (2010).
64. Crescente, J. M., Zavallo, D., Helguera, M. & Vanzetti, L. S. MITE Tracker: an accurate approach to identify miniature inverted-repeat transposable elements in large genomes. *BMC Bioinforma.* **19**, 348 (2018).
65. Xiong, W., He, L., Lai, J., Dooner, H. K. & Du, C. HelitronScanner uncovers a large overlooked cache of Helitron transposons in many plant genomes. *Proc. Natl Acad. Sci. USA* **111**, 10263–10268 (2014).
66. Wicker, T. et al. A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* **8**, 973–982 (2007).
67. Melters, D. P. et al. Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. *Genome Biol.* **14**, R10 (2013).
68. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
69. Tamura, K. et al. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* **28**, 2731–2739 (2011).
70. Campbell, M. S. et al. MAKER-P: a tool kit for the rapid creation, management, and quality control of plant genome annotations. *Plant Physiol.* **164**, 513–524 (2014).
71. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
72. Pertea, M. et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).
73. Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M. & Bairoch, A. in *Plant Bioinformatics* (Springer, 2007).
74. Korfi, I. Gene finding in novel genomes. *BMC Bioinforma.* **5**, 59 (2004).
75. Stanke, M. & Waack, S. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **19**, ii215–ii225 (2003).
76. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
77. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).
78. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
79. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
80. Suyama, M., Torrents, D. & Bork, P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* **34**, W609–W612 (2006).

Acknowledgements

We are indebted to Tsegaye Dabi at the Salk Institute for Biological Studies for introducing us to this amazing plant, and for inspiring generations of plant biologists. We thank Elliott Meer for assistance with PacBio sequencing, and the Monsanto Genomics Team (Randy Kerstetter, Mitch Sudkamp, Phil Latreille, Zijin Du, and Joe Zhou) for full-length sequenced fosmids. We thank James Schnable for his helpful comments and suggestions on the manuscript. This work is supported by funding from the National Science Foundation (MCB-1817347 to R.V.), Department of Energy (DE-SC0012639 to T.C.M. and T.P.M.), and partial support from the Bill & Melinda Gates Foundation (T.C.M. and D.B.). This work is supported by the USDA National Institute of Food and Agriculture Hatch project to R.V.

Author contributions

R.V. and T.P.M. conceived this project and coordinated research activities. R.V., C.M.W., S.R.C., G.H., D.B., J.M., M.E.S., T.C.M., J.L.B. and T.P.M. processed genome sequencing, genetic map, and expression data. R.V., C.M.W., X.W., J.P., A.E.Y., H.W., P.P.E., J.L.B. and T.P.M. analyzed genome-scale and expression data. R.V., J.L.B. and T.P.M. wrote the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41467-020-14724-z>.

Correspondence and requests for materials should be addressed to R.V. or T.P.M.

Peer review information *Nature Communications* thanks Andrea Zuccolo, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020