nature portfolio

https://doi.org/10.1038/s41586-026-10229-9

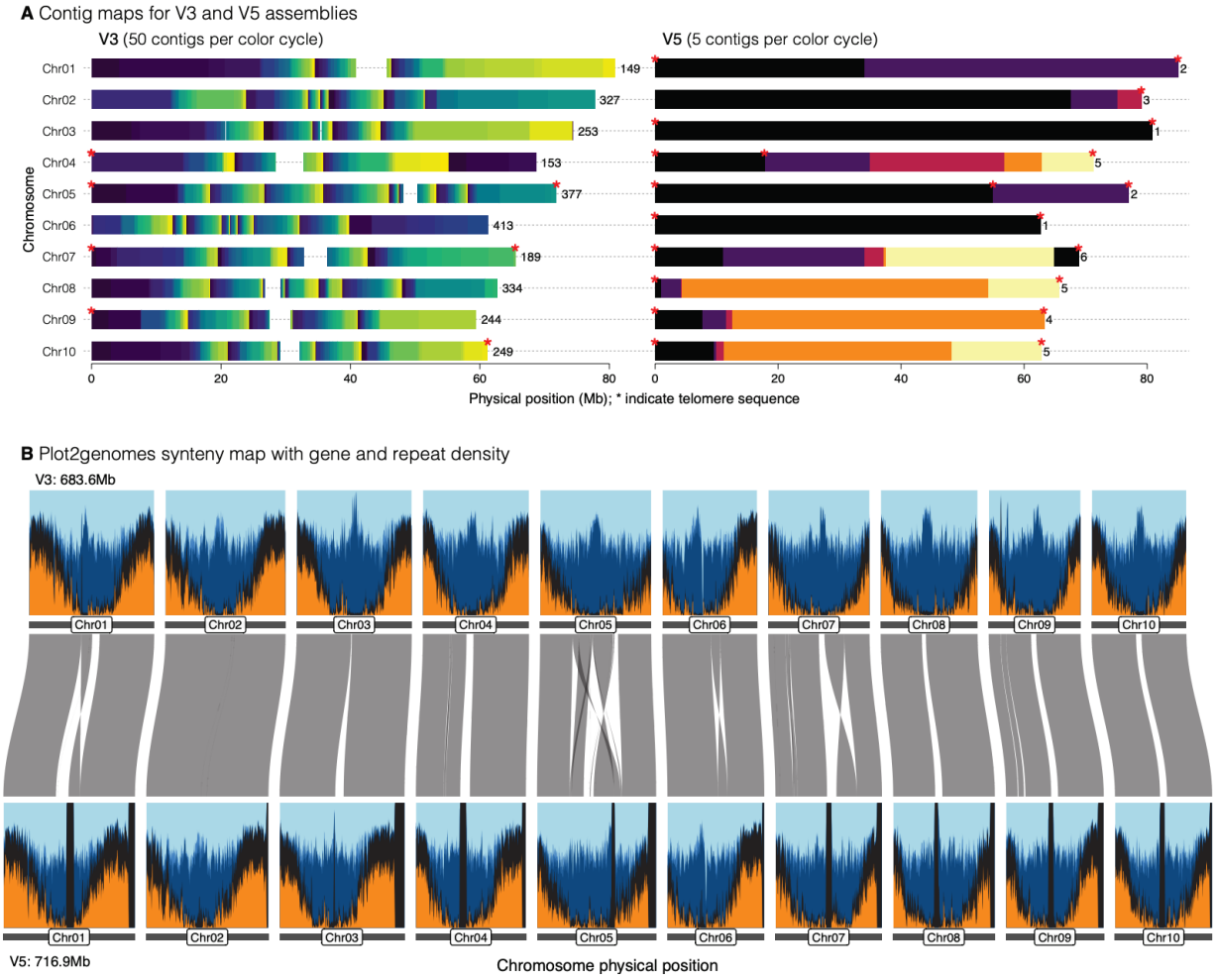
Supplementary information

A sorghum pangenome reference improves global crop trait discovery

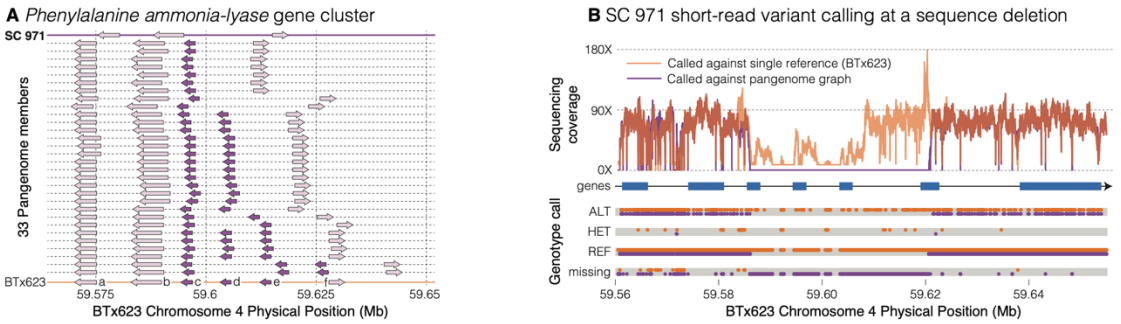
In the format provided by the authors and unedited

Supplementary information for: *A sorghum pangenome reference improves global crop trait discovery*

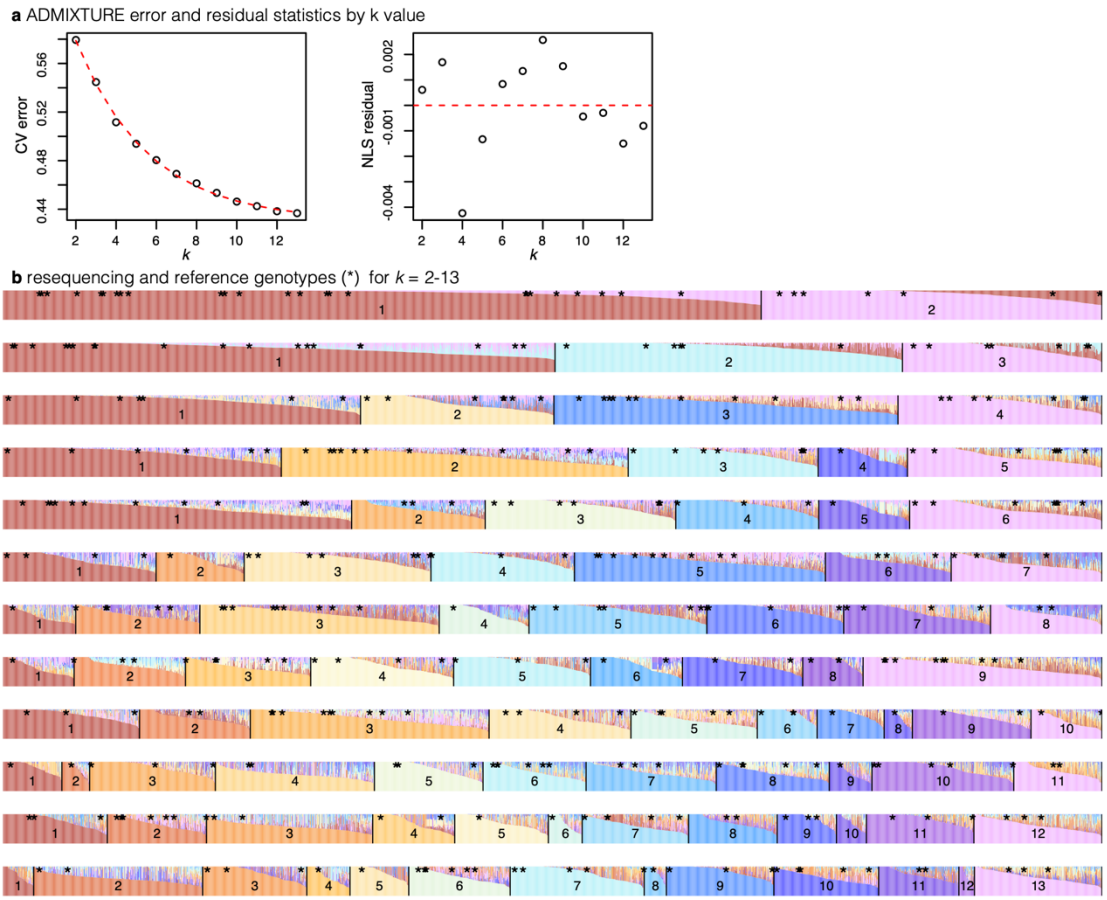
SUPPLEMENTARY FIGURES



Supplementary Figure 1 | Genome assembly and annotation comparison between V3 and V5 BTx623 resources. A Contig maps showing the position of gaps and telomeres in V3 (left) and V5 (right) illustrate an approximate 100-fold increase in assembly quality between the versions. Red *s indicate the presence of dense blocks of putative telomeric repeats. B Synteny map accompanied by gene (orange) and repeat density (blues: dark = T3, light = other; 100kb-overlapping 500kb sliding windows) show the position of structural variants and increased repeat representation between the two versions.



Supplementary Figure 2 | Comparison of graph and single-reference genotyping approaches at a known presence-absence variation site. A Presence-absence variation at a phenylalanine ammonia lyase on chromosome 4 across the 33 pangenome members. Arrows indicate the presence and orientation of the bounding genes that all are single copy (lavender) or the CNV variable phenylalanine ammonia lyase gene members (dark purple). From left to right the six gene models in BTx623 are: a: Sobic.004G220300, b: Sobic.004G220400, c: Sobic.004G220500, d: Sobic.004G220600, e: Sobic.004G220650, and f: Sobic.004G220700. PI numbers (or genome IDs) from bottom to top: BTx623, PI656057, PI597980, PI656044, PI656027, PI565121, PI276837, PI513676, PI569459, RTx430, Wray, PI329301, PI276816, PI329501, PI656023, PI180348, PI534133, PI655981, PI660557, PI533766, PI660565, PI585966, PI570071, PI156178, PI660563, PI656031, PI656015, PI655988, PI576434, PI154844, BTx642, PI656050, PI656111. Positions are provided in the primary reference in the pangenome graph, BTx623, coordinate system. Positions of genes in other genomes are approximate. B Comparison of depth (sequencing coverage) and genotype calls between linear (orange) and graph (purple) alignments of the phenylalanine ammonia lyase array on sorghum chromosome 4.



Supplementary Figure 3 | Choosing the optimal subpopulation k for ADMIXTURE. A 'CV error' and non-linear regression (NLS) statistic for *k* = 2-13. B Ancestry proportions for unique genotypes of the resequencing libraries (*n* = 1,984) represented as barplots for *k* = 2-13. The bars (resequencing libraries) are horizontally ordered by subpopulation ID, then by ancestry therein. Subpopulation IDs are not equivalent between *k*-values, so order from one *k* to the next is not consistent. Polishing libraires of the 33 pangenome reference members are flagged (*) above the barplot for each *k* value.

SUPPLEMENTARY TABLES

Supplementary Table 1 | Genome assembly statistics. Combined statistics, genome IDs, EMBL accession IDs, and common names for the 31 new reference genomes presented here and *two previously published genomes used in the pangenome resource. Full methodological descriptions can be found in Supplementary Note 2.

Genome Name	Common Name	Genome (Mb)	Contig N50 (Mb)	Genome in chrs(%)	Annotation BUSCO (%)	PacBio Coverage	STUDY ID	SAMPLE ID
BTx623	BTx623	719.9	50.7	99.59	99.7	61.47	PRJEB107652	ERS28674301
BTx642*	BTx642	680.1	12.1	99.45	99.4	52.36	PRJEB107621	ERS28674303
PI_154844	GRASSL	701.2	28.4	99.02	99.4	49.4	PRJEB107303	ERS28647261
PI_156178	MN 2014	681.1	20	99.62	99.4	56.16	PRJEB107622	ERS28674304
PI_180348	Juar (IS 12876)	690.6	15.1	99.33	99.6	89.12	PRJEB107623	ERS28674305
PI_276816	IS 12646	699.3	21.4	98.85	99.5	77.85	PRJEB107624	ERS28674306
PI_276837	IS 12661	693.7	18.7	98.94	99.7	54.36	PRJEB107625	ERS28674307
PI_329301	IS 11059	683.3	15.7	98.93	99.7	48.99	PRJEB107626	ERS28674308
PI_329501	IS 11257	765.8	2.6	89.84	99.2	64.83	PRJEB107627	ERS28674309
PI_513676	Matche-Dakomia (IS 20380)	688.8	29.3	99.06	99.4	61.76	PRJEB107628	ERS28674310
PI_533766	SC 265	687.8	22.4	99.35	99.3	82.82	PRJEB107629	ERS28674311
PI_534133	SC 35	688.6	15.4	98.86	99.4	70.32	PRJEB107630	ERS28674312
PI_565121	MACIA	689.1	26.1	99.86	99.7	51.46	PRJEB107631	ERS28674313
PI_569459	LUWAIAH (IS 19523)	716.2	5.3	95.01	99.7	55.74	PRJEB107632	ERS28674314
PI_570071	IS 22849	667.5	4.3	98.33	98.2	49.38	PRJEB107633	ERS28674315
PI_576434	SC 1103	688.7	20.1	99.04	99.5	63.4	PRJEB107634	ERS28674316
PI_585966	DJOFELA (IS 26395)	681.7	9.5	98.72	97.7	53.59	PRJEB107635	ERS28674317
PI_597980	SC 1345 (CSM-90)	692.1	23	99.15	99.6	63.87	PRJEB107636	ERS28674318
PI_655981	CSM-63	692.1	22	99.38	99.5	52.37	PRJEB107637	ERS28674319
PI_655988	COMBINE KAFIR-60	696.8	24.3	99.33	99.7	53.03	PRJEB107638	ERS28674320
PI_656015	Ajabsido	689.4	22.6	98.68	99.6	55.71	PRJEB107639	ERS28674321
PI_656023	Segaolane	696.2	27.9	99.28	99.4	88.11	PRJEB107640	ERS28674322
PI_656027	SRN39	696.7	18.3	99.32	99.7	62.89	PRJEB107641	ERS28674323
PI_656031	IRAT204	741.4	20.7	92.8	99.6	51.19	PRJEB107642	ERS28674324
PI_656044	Kuyuma	686.5	14.1	98.95	99.6	44.34	PRJEB107643	ERS28674325
PI_656050	Mota Maradi	698.8	20.9	99.61	99.5	72.21	PRJEB107644	ERS28674326
PI_656057	P898012	694.7	17.8	99.1	99.4	67.7	PRJEB107645	ERS28674327
PI_656111	SC 971	690.9	12.1	98.63	99.4	62.89	PRJEB107646	ERS28674328
PI_660557	IS 6705	690.1	21.5	99.04	99.3	49.95	PRJEB107647	ERS28674329
PI_660563	IS 7173	698.1	27.7	99.05	99.4	81.81	PRJEB107648	ERS28674330
PI_660565	Kamjin (IS 3620)	707.6	25.4	99	99.4	87.76	PRJEB107620	ERS28674331
RTx430*	RTx430	677.7	4	98.09	99.2	80.41	PRJEB107649	ERS28674332
Wray	Wray (PI597980)	736.8	63.9	98.72	99.4	31.97	PRJEB107650	ERS28674333

Supplementary Table 2 | Genome annotation sequencing and data deposition information. NCBI deposition and Illumina RNA-seq and PacBio Iso-Seq data summaries used for genome annotation of 31 new reference genomes presented in this study.

Genome Name	Illumina RNA-seq read pairs (Billion)	PacBio Iso-Seq CCS reads (Million)	BioSample	Bioproject
BTx623	4.0	19.73	SAMN02953738	PRJNA13876
Wray	4.0	24	SAMN49011330	PRJNA1275079
PI_329301	0.79	7.65	SAMN49013189	PRJNA1275246
PI_156178	0.76	9.8	SAMN10864443	PRJNA518927
PI_180348	0.9	8.61	SAMN49012613	PRJNA1275204
PI_276837	0.73	-	SAMN49724291	PRJNA1284433
PI_569459	0.27	9.94	SAMN10864442	PRJNA518925
PI_570071	1.83	9.95	SAMN10864444	PRJNA518971
PI_656031	0.7	-	SAMN49724203	PRJNA1284421
PI_656044	0.71	6.04	SAMN49011439	PRJNA1275156
PI_154844	0.67	-	SAMN49724292	PRJNA1284412
PI_276816	1.45	-	SAMN49724163	PRJNA1284427
PI_329501	0.95	7.72	SAMN10864446	PRJNA518970
PI_513676	0.81	-	SAMN49724256	PRJNA1284432
PI_533766	0.74	-	SAMN49724255	PRJNA1284431
PI_534133	1.19	-	SAMN49724249	PRJNA1284429
PI_565121	0.84	-	SAMN49724201	PRJNA1284430
PI_576434	0.61	-	SAMN49724202	PRJNA1284417
PI_585966	0.59	6.9	SAMN49011440	PRJNA1275247
PI_597980	0.9	-	SAMN49724220	PRJNA1284426
PI_655981	0.79	-	SAMN49724164	PRJNA1284424
PI_655988	0.74	5.72	SAMN10864445	PRJNA518926
PI_656015	0.76	-	SAMN49724218	PRJNA1284425
PI_656023	0.86	-	SAMN49724217	PRJNA1284423
PI_656027	0.88	-	SAMN49724199	PRJNA1284422
PI_656050	0.71	-	SAMN49724200	PRJNA1284420
PI_656057	0.8	-	SAMN49724219	PRJNA1284419
PI_656111	0.79	-	SAMN49724162	PRJNA1284415
PI_660557	0.78	-	SAMN49724161	PRJNA1284414
PI_660563	0.92	-	SAMN49724160	PRJNA1284413
PI_660565	1.47	-	SAMN49724159	PRJNA1284416

Supplementary Table 3 | Transcription factor binding sites in the Dhurrin BGC. Predicted transcription factor binding sites and their overlap with regions of accessible chromatin for select intergenic sites strongly associated with the dhurrin seedling content phenotype. The two sites that do not have a P-value reported were not included in GWAS due to high missingness.

Site (BTx623 coordinates)	Location	log_{10}(P-wald) Seedling dhurrin content	TF Recognition Site(s) Identified through PlantPAN (Y/N)	TF ID or Motif name	Accessible chromatin region overlap with site identified through SorghumBase (Y/N)
INDEL Chr01:1,059,176bp	Intergenic between Sobic.001G012100 (BGC 5' border gene) and *CYP71E1*	-5.76581	Yes	AT3G14230 (RAP2.2), AT5G65410 (AtHB25), AT1G23420 (INO), AT4G29940 (PRHA), AT1G75240 (AtHB33), AT1G09030 (NF-YB4), AT1G75520 (SRS5)	No
SNP Chr01:1,079,053bp	Intergenic between *CYP71E1* and *CYP79A1*	-3.32338	Yes	AT1G23380 (KNAT6), AT1G77920 (TGA7)	No
SNP Chr01:1,094,821bp	Intergenic between *CYP71E1* and *CYP79A1*		No	NA	No
SNP Chr01:1,149,356bp	Intergenic between *CYP71E1* and *CYP79A1*		Yes	AT2G42540 (AtCOR15A)	No
INDEL Chr01:1,185,567bp	Intergenic between *UGT85B1* and *GST1*	-4.876768	Yes	AT3G19290 (ABF4), AT4G34000 (ABF3)	Yes

Supplementary Table 4 | Summary of UAV flights and data collection. For each mission we report date, crop growth stage, altitude above ground level (AGL), ground-sampling distance (GSD; cm/pixel), local solar time imaging window, reflectance calibration (panel/irradiance) performed before and after flights, radiometric correction, ground control points (GCPs), photogrammetry software and version, and sensors used. Abbreviations: DAS, days after sowing; GSD, ground-sampling distance; AGL, above ground level; R, red; G, green; B, blue; RE, red edge; NIR, near infrared; RGB, three-band visible camera; MS, multispectral camera; RTK, real-time kinematic; GCPs, ground control points; ODM, OpenDroneMap; Reflect. Calib., reflectance panel/irradiance calibration; Rad. Cor., radiometric correction; Flo-Mat, flowering to maturity. When two GSD values are listed (e.g., 2.18 & 3.6), they correspond to RGB and MS mosaics, respectively. Times are local (09:00–13:00). Approximate altitudes are noted as "~50 m AGL".

Flight #	Date	DAS	Site	Growth	Altitude	GSD (cm/pixel)	Time	Reflect. Calib.	Rad. Cor.	GCPs	Software	Sensors
1	6/9/23	0	MO	Bare soil	~50 m AGL	2.18 & 3.6	9AM-1PM	bef. & aft. Flight	Yes	RTK targets	ODM 2.9.1	RGB & MS
2	6/15/23	0	MO	Bare soil	~50 m AGL	2.18 & 3.6	9AM-1PM	bef. & aft. Flight	Yes	RTK targets	ODM 2.9.1	RGB & MS
3	6/21/23	5	MO	Emergence	~50 m AGL	2.18 & 3.6	9AM-1PM	bef. & aft. Flight	Yes	RTK targets	ODM 2.9.1	RGB & MS
4	6/29/23	13	MO	Seedling	~50 m AGL	2.18 & 3.6	9AM-1PM	bef. & aft. Flight	Yes	RTK targets	ODM 2.9.1	RGB & MS
5	7/7/23	21	MO	Vegetative	~50 m AGL	2.18 & 3.6	9AM-1PM	bef. & aft. Flight	Yes	RTK targets	ODM 2.9.1	RGB & MS
6	7/12/23	26	MO	Vegetative	~50 m AGL	2.18 & 3.6	9AM-1PM	bef. & aft. Flight	Yes	RTK targets	ODM 2.9.1	RGB & MS
7	7/17/23	31	MO	Vegetative	~50 m AGL	2.18 & 3.6	9AM-1PM	bef. & aft. Flight	Yes	RTK targets	ODM 2.9.1	RGB & MS
8	7/27/23	41	MO	Booting	~50 m AGL	2.18 & 3.6	9AM-1PM	bef. & aft. Flight	Yes	RTK targets	ODM 2.9.1	RGB & MS
9	8/4/23	49	MO	Booting	~50 m AGL	2.18 & 3.6	9AM-1PM	bef. & aft. Flight	Yes	RTK targets	ODM 2.9.1	RGB & MS
10	8/10/23	55	MO	Flowering	~50 m AGL	2.18 & 3.6	9AM-1PM	bef. & aft. Flight	Yes	RTK targets	ODM 2.9.1	RGB & MS
11	8/17/23	62	MO	Flowering	~50 m AGL	2.18 & 3.6	9AM-1PM	bef. & aft. Flight	Yes	RTK targets	ODM 2.9.1	RGB & MS
12	8/22/23	67	MO	Flo-Mat	~50 m AGL	2.18 & 3.6	9AM-1PM	bef. & aft. Flight	Yes	RTK targets	ODM 2.9.1	RGB & MS
13	8/31/23	76	MO	Flo-Mat	~50 m AGL	2.18 & 3.6	9AM-1PM	bef. & aft. Flight	Yes	RTK targets	ODM 2.9.1	RGB & MS
14	7/2/24	26	MO	Vegetative	~50 m AGL	1.45	9AM-1PM	bef. & aft. Flight	Yes	RTK targets	ODM 2.9.1	RGB
15	8/1/24	56	MO	Flowering	~50 m AGL	1.45	9AM-1PM	bef. & aft. Flight	Yes	RTK targets	ODM 2.9.1	RGB
16	9/4/24	90	MO	Maturity	~50 m AGL	1.45	9AM-1PM	bef. & aft. Flight	Yes	RTK targets	ODM 2.9.1	RGB
17	6/21/24	15	MO	Vegetative	~50 m AGL	2.29	9AM-1PM	bef. & aft. Flight	Yes	RTK targets	ODM 2.9.1	MS
18	7/10/24	34	MO	Vegetative	~50 m AGL	2.29	9AM-1PM	bef. & aft. Flight	Yes	RTK targets	ODM 2.9.1	MS
19	7/19/24	43	MO	Booting	~50 m AGL	2.29	9AM-1PM	bef. & aft. Flight	Yes	RTK targets	ODM 2.9.1	MS
20	8/5/24	60	MO	Flowering	~50 m AGL	2.29	9AM-1PM	bef. & aft. Flight	Yes	RTK targets	ODM 2.9.1	MS
21	8/19/24	74	MO	Flowering	~50 m AGL	2.29	9AM-1PM	bef. & aft. Flight	Yes	RTK targets	ODM 2.9.1	MS
22	8/30/24	85	MO	Maturity	~50 m AGL	2.29	9AM-1PM	bef. & aft. Flight	Yes	RTK targets	ODM 2.9.1	MS
23	9/12/24	98	MO	Maturity	~50 m AGL	2.29	9AM-1PM	bef. & aft. Flight	Yes	RTK targets	ODM 2.9.1	MS
24	9/25/24	111	MO	Maturity	~50 m AGL	2.29	9AM-1PM	bef. & aft. Flight	Yes	RTK targets	ODM 2.9.1	MS
25	7/19/23	34	SC	Vegetative	~50 m AGL	2.3	9AM-1PM	bef. & aft. Flight	Yes	RTK targets	ODM 2.9.2	MS
26	9/21/23	96	SC	Maturity	~50 m AGL	2.3	9AM-1PM	bef. & aft. Flight	Yes	RTK targets	ODM 2.9.3	MS
27	10/7/24		SC	Maturity	~50 m AGL	2.3	9AM-1PM	bef. & aft. Flight	Yes	RTK targets	ODM 2.9.4	MS

SUPPLEMENTARY NOTES

Supplementary Note 1 | Description of the internal BTx623 'V4' release

V4 was never released publicly because of the upcoming availability of PacBio CLR technology. Below we provide the release notes describing the differences between V3 and V4. This release corrects an omission in the V3 release. There were a total of 4 scaffolds (~23Mb) that were inadvertently left in the release. These are scaffolds that were integrated into the chromosomes using the new mapping resources but were not identified as duplicates during the final analysis. They are super_10, super_11, super_12, and super_610. They have been removed in this release. Also, all numbering is the same as the V3 release. V4 represents the second improved *Sorghum bicolor* release which includes ~351 Mb of finished sorghum sequence. These regions were finished by dividing the gene space into ~1Mb overlapping pieces. Each region was manually inspected and then finished using a variety of technologies including Sanger (primer walks on subclones and fosmid templates, transposon sequencing on subclone templates, shotgun sequencing and finishing of complete fosmid and BAC clones that were then integrated back into the genome pieces) 454 (small insert shatter libraries) and Illumina (small insert shatter libraries). Following completion each assembly was validated by an independent quality assessment. This included a visual examination of subclone paired ends and repeat structures and validation of any remaining lower quality regions and regions with high quality base pair discrepancies. These improved sequences have an estimated error rate of less than one error in 100,000 base pairs. Integration of the finished regions began by aligning the regions to the existing V1.0 assembly and assigning regions to a chromosome. Orientation of individual contigs was adjusted to get all the regions on the same strand. Overlapping regions were first merged together to produce megaRegions. Ends of the megaRegions were then placed on the genome to identify the start and end of the insertion. Sequence and quality scores were then integrated into the assembly, and statistics on the overlap region were obtained (number of gaps, total Ns, etc.). The integrated assembly was screened for *E. coli* and adapter sequences, and a total of 10,339 bp was identified as foreign and excised from the assembly. A total of 349 clones were integrated representing 344.4 MB of sequence. 4,426 gaps were closed, and a total of 4.96 MB of sequence was added to the assembly. Overall contiguity (contig N50) increased by a factor of 5.8x from 204.5 KB to 1.2 MB.

Supplementary Note 2 | Full genome assembly methods for each new reference genome

Note BTx642 and RTx430 have been published previously.

BTx623 (V5) | The BTx623 assembly was generated by assembling the PACBIO CLR (61.47x coverage, average read size 11,008 bp) reads using the MECAT (Xiao et al., 2017) assembler and subsequently polished using RACON (Vaser et al., 2017). This produced an initial assembly consisting of 738.7 Mb sequence assembled and a contig N50 of 14.5 Mb. A combination of syntenic markers from BTx642 and primary annotated genes from BTx623 were used to identify misjoins in the assembly. Syntenic markers consisted of a total of 32,400 unique, non-repetitive, non-overlapping 1.0 Kb sequences generated using the version 1.0 S. bicolor BTx642 genome release. The gene set consisted of a subset of 12,641 BTx642 annotated primary genes from the V1 annotation. Both the syntenic markers and the primary gene set were aligned to the polished assembly and a total of 9 misjoins were identified in the assembly. Contigs were then ordered, oriented, and assembled into chromosomes using the combination of syntenic markers and genes. Contigs terminating in significant telomeric sequence were identified using the (TTTAGGG)n repeat, and care was taken to make sure that they were properly oriented in the release assembly. A total of 32 joins were applied to the assembly to form the final assembly consisting of 10 chromosomes. Each chromosome join is padded with 10,000 Ns. Chromosomes were numbered using the V1.0 S. bicolor BTx642 assembly (Cole et al., 2025) available from Phytozome (Goodstein et al., 2012). The remaining scaffolds were screened against bacterial proteins, organelle sequences, GenBank nr and removed if found to be a contaminant. Chloroplast and mitochondrial genomes were generated using the OatK pipeline (Zhou et al., 2024). Homozygous SNPs and INDELs were corrected in the release using 53.5x of the Illumina fragment reads (2x150, 400bp insert) by aligning the reads using bwa mem (Li & Durbin, 2009) and identifying homozygous SNPs and INDELs with the GATK's UnifiedGenotyper tool (McKenna et al., 2010). A total of 254 homozygous snps and 2,443 homozygous indels were corrected in the consensus sequence. The final release consisted of 719.9 Mb of sequence with a contig N50 of 50.7 Mb.

BTx642 | The BTx642 assembly was generated by assembling the PACBIO CLR (52.36x coverage, average read size 7,024 bp) reads using the MECAT (Xiao et al., 2017) assembler and subsequently polished using ARROW (Chin et al., 2013). This produced an initial assembly consisting of 714.9 Mb sequence assembled and a contig N50 of 8.3 Mb. A combination of syntenic markers from BTx642 and primary annotated genes from BTx623 were used to identify misjoins in the assembly. Syntenic markers consisted of a total of 32,400 unique, non-repetitive, non-overlapping 1.0 Kb sequences generated using the version 1.0 S. bicolor BTx642 genome release. The gene set consisted of a subset of 12,641 BTx642 annotated primary genes from the V1 annotation. Both the syntenic markers and the primary gene set were aligned to the polished assembly and 16 misjoins were identified in the assembly. Contigs were then ordered, oriented, and assembled into chromosomes using the combination of syntenic markers and genes. Contigs terminating in significant telomeric sequence were identified using the (TTTAGGG)n repeat, and care was taken to make sure that they were properly oriented in the release assembly. A total of 155 joins were applied to the assembly to form the final assembly consisting of 10 chromosomes. Each chromosome join is padded with 10,000 Ns. Chromosomes were numbered using the V5.0 S. bicolor BTx623 assembly available from Phytozome (Goodstein et al., 2012). The remaining scaffolds were screened against bacterial proteins, organelle sequences, GenBank nr and removed if found to be a contaminant. Chloroplast and mitochondrial genomes were generated using the OatK pipeline (Zhou et al., 2024). Homozygous SNPs and INDELs were corrected in the release using 55.4x of the Illumina fragment reads (2x150, 400bp insert) by aligning the reads using bwa mem (Li, 2013) and identifying homozygous SNPs and INDELs with the GATK's UnifiedGenotyper tool (McKenna et al., 2010). A total of 945 homozygous snps and 6,347 homozygous indels were corrected in the consensus sequence. The final release consisted of 680.1 Mb of sequence with a contig N50 of 12.1 Mb

Wray (PI597980) | The Wray assembly was generated by assembling the PACBIO HiFi (31.97x coverage, average read size 21,161 bp) reads using the HiFiAsm assembler (Cheng et al., 2021) assembler and subsequently polished using RACON (Vasser et al., 2017). This produced an initial assembly consisting of 765.6 Mb sequence assembled and a contig N50 of 65.8 Mb. A combination of syntenic markers from BTx642 and primary annotated genes from BTx623 were used to identify misjoins in the assembly. Syntenic markers consisted of a total of 32,400 unique, non-repetitive, non-overlapping 1.0 Kb sequences generated using the version 1.0 S. bicolor BTx642 genome release. The gene set consisted of a subset of 12,641 BTx642 annotated primary genes from the V1 annotation. Both the syntenic markers and the primary gene set were aligned to the polished assembly and no misjoins were identified in the assembly. Contigs were then ordered, oriented, and assembled into chromosomes using the combination of syntenic markers and genes. Contigs terminating in significant telomeric sequence were identified using the (TTTAGGG)n repeat, and care was taken to make sure that they were properly oriented in the release assembly. A total of 7 joins were applied to the assembly to form the final assembly consisting of 10 chromosomes. Each chromosome join is padded with 10,000 Ns. Chromosomes were numbered using the V1.0 S. bicolor BTx642 assembly available from Phytozome (Goodstein et al., 2012). The remaining scaffolds were screened against bacterial proteins, organelle sequences, GenBank nr and removed if found to be a contaminant. Chloroplast and mitochondrial genomes were generated using the OatK pipeline (Zhou et al., 2024). Homozygous SNPs and INDELs were corrected in the release using 52.0x of the Illumina fragment reads (2x150, 400bp insert) by aligning the reads using bwa mem (Li, 2013) and identifying homozygous SNPs and INDELs with the GATK's UnifiedGenotyper tool (McKenna et al., 2010). A total of 31 homozygous snps and 2,019 homozygous indels were corrected in the consensus sequence. The final release consisted of 736.7 Mb of sequence with a contig N50 of 63.9 Mb.

RTx430 | The RTx430 assembly was generated by assembling the PACBIO CLR (80.41x coverage, average read size 8,091 bp) reads using the MECAT (Xiao et al., 2017) assembler and subsequently polished using ARROW (Chin et al., 2013). This produced an initial assembly consisting of 704.9 Mb sequence assembled and a contig N50 of 2.3 Mb. A combination of syntenic markers from BTx642 and primary annotated genes from BTx623 were used to identify misjoins in the assembly. Syntenic markers consisted of a total of 32,400 unique, non-repetitive, non-overlapping 1.0 Kb sequences generated using the version 1.0 S. bicolor BTx642 genome release. The gene set consisted of a subset of 12,641 BTx642 annotated primary genes from the V1 annotation. Both the syntenic markers and the primary gene set were aligned to the polished assembly and a total of 32 misjoins were identified in the assembly. Contigs were then ordered, oriented, and assembled into chromosomes using the combination of syntenic markers and genes. Contigs terminating in significant telomeric sequence were identified using the (TTTAGGG)n repeat, and care was taken to make sure that they were properly oriented in the release assembly. A total of 553 joins were applied to the assembly to form the final assembly consisting of 10 chromosomes. Each chromosome join is padded with 10,000 Ns. Chromosomes were numbered using the V1.0 S. bicolor BTx642 assembly available from Phytozome (Goodstein et al., 2012). The remaining scaffolds were screened against bacterial proteins, organelle sequences, GenBank nr and removed if found to be a contaminant. Chloroplast and mitochondrial genomes were generated using the OatK pipeline (Zhou et al., 2024). Homozygous SNPs and INDELs were corrected in the release using 60.0x of the Illumina fragment reads (2x150, 400bp insert) by aligning the reads using bwa mem (Li, 2013) and identifying homozygous SNPs and INDELs with the GATK's UnifiedGenotyper tool (McKenna et al., 2010). A total of 1,617 homozygous snps and 5,954 homozygous indels were corrected in the consensus sequence. The final release consisted of 677.7 Mb of sequence with a contig N50 of 4.0 Mb.

PI_329301 (IS 11059) | The PI_329301 (IS 11059) assembly was generated by assembling the PACBIO CLR (49.88x coverage, average read size 9,047 bp) reads using the CANU (Koren et al., 2017) assembler and subsequently polished using ARROW (Chin et al., 2013). This produced an initial assembly consisting of 713.2 Mb sequence assembled and a contig N50 of 10.1 Mb. A combination of syntenic markers from BTx642 and primary annotated genes from BTx623 were used to identify misjoins in the assembly. Syntenic markers consisted of a total of 32,400 unique, non-repetitive, non-overlapping 1.0 Kb sequences generated using the version 1.0 S. bicolor BTx642 genome release. The gene set consisted of a subset of 12,641 BTx642 annotated primary genes from the V1 annotation. Both the syntenic markers and the primary gene set were aligned to the polished assembly and a total of 2 misjoins were identified in the assembly. Contigs were then ordered, oriented, and assembled into chromosomes using the combination of syntenic markers and genes. Contigs terminating in significant telomeric sequence were identified using the (TTTAGGG)n repeat, and care was taken to make sure that they were properly oriented in the release assembly. A total of 196 joins were applied to the assembly to form the final assembly consisting of 10 chromosomes. Each chromosome join is padded with 10,000 Ns. Chromosomes were numbered using the V1.0 S. bicolor BTx642 assembly available from Phytozome (Goodstein et al., 2012). The remaining scaffolds were screened against bacterial proteins, organelle sequences, GenBank nr and removed if found to be a contaminant. Chloroplast and mitochondrial genomes were generated using the OatK pipeline (Zhou et al., 2024). Homozygous SNPs and INDELs were corrected in the release using 60.8x of the Illumina fragment reads (2x150, 400bp insert) by aligning the reads using bwa mem (Li, 2013) and identifying homozygous SNPs and INDELs with the GATK's UnifiedGenotyper tool (McKenna et al., 2010). A total of 268 homozygous snps and 6,278 homozygous indels were corrected in the consensus sequence. The final release consisted of 681.8 Mb of sequence with a contig N50 of 15.7 Mb.

PI_156178 (MN 2014) | The PI_156178 (MN 2014) assembly was generated by assembling the PACBIO CLR (56.16x coverage, average read size 9,332 bp) reads using the MECAT (Xiao et al., 2017) assembler and subsequently polished using ARROW (Chin et al., 2013). This produced an initial assembly consisting of 708.8 Mb sequence assembled and a contig N50 of 13.2 Mb. A combination of syntenic markers from BTx642 and primary annotated genes from BTx623 were used to identify misjoins in the assembly. Syntenic markers consisted of a total of 32,400 unique, non-repetitive, non-overlapping 1.0 Kb sequences generated using the version 1.0 S. bicolor BTx642 genome release. The gene set consisted of a subset of 12,641 BTx642 annotated primary genes from the V1 annotation. Both the syntenic markers and the primary gene set were aligned to the polished assembly and a total of 9 misjoins were identified in the assembly. Contigs were then ordered, oriented, and assembled into chromosomes using the combination of syntenic markers and genes. Contigs terminating in significant telomeric sequence were identified using the (TTTAGGG)n repeat, and care was taken to make sure that they were properly oriented in the release assembly. A total of 140 joins were applied to the assembly to form the final assembly consisting of 10 chromosomes. Each chromosome join is padded with 10,000 Ns. Chromosomes were numbered using the V1.0 S. bicolor BTx642 assembly available from Phytozome (Goodstein et al., 2012). The remaining scaffolds were screened against bacterial proteins, organelle sequences, GenBank nr and removed if found to be a contaminant. Chloroplast and mitochondrial genomes were generated using the OatK pipeline (Zhou et al., 2024). Homozygous SNPs and INDELs were corrected in the release using 37.1x of the Illumina fragment reads (2x150, 400bp insert) by aligning the reads using bwa mem (Li, 2013) and identifying homozygous SNPs and INDELs with the GATK's UnifiedGenotyper tool (McKenna et al., 2010). A total of 129 homozygous snps and 2,734 homozygous indels were corrected in the consensus sequence. The final release consisted of 681.1 Mb of sequence with a contig N50 of 20.0 Mb.

PI_180348 (Juar / IS 12876) | The PI_180348 (Juar / IS 12876) assembly was generated by assembling the PACBIO CLR (89.12x coverage, average read size 9,415 bp) reads using the CANU (Koren et al., 2017) assembler and subsequently polished using ARROW (Chin et al., 2013). This produced an initial assembly consisting of 726.6 Mb sequence assembled and a contig N50 of 11.2 Mb. A combination of syntenic markers from BTx642 and primary annotated genes from BTx623 were used to identify misjoins in the assembly. Syntenic markers consisted of a total of 32,400 unique, non-repetitive, non-overlapping 1.0 Kb sequences generated using the version 1.0 S. bicolor BTx642 genome release. The gene set consisted of a subset of 12,641 BTx642 annotated primary genes from the V1 annotation. Both the syntenic markers and the primary gene set were aligned to the polished assembly and no misjoins were identified in the assembly. Contigs were then ordered, oriented, and assembled into chromosomes using the combination of syntenic markers and genes. Contigs terminating in significant telomeric sequence were identified using the (TTTAGGG)n repeat, and care was taken to make sure that they were properly oriented in the release assembly. A total of 165 joins were applied to the assembly to form the final assembly consisting of 10 chromosomes. Each chromosome join is padded with 10,000 Ns. Chromosomes were numbered using the V1.0 S. bicolor BTx642 assembly available from Phytozome (Goodstein et al., 2012). The remaining scaffolds were screened against bacterial proteins, organelle sequences, GenBank nr and removed if found to be a contaminant. Chloroplast and mitochondrial genomes were generated using the OatK pipeline (Zhou et al., 2024). Homozygous SNPs and INDELs were corrected in the release using 31.4x of the Illumina fragment reads (2x150, 400bp insert) by aligning the reads using bwa mem (Li, 2013) and identifying homozygous SNPs and INDELs with the GATK's UnifiedGenotyper tool (McKenna et al., 2010). A total of 247 homozygous snps and 15,144 homozygous indels were corrected in the consensus sequence. The final release consisted of 690.6 Mb of sequence with a contig N50 of 15.1 Mb.

PI_276837 (IS 12661) | The PI_276837 (IS 12661) assembly was generated by assembling the PACBIO CLR (54.36x coverage, average read size 11,931 bp) reads using the CANU (Koren et al., 2017) assembler and subsequently polished using ARROW (Chin et al., 2013). This produced an initial assembly consisting of 737.9 Mb sequence assembled and a contig N50 of 11.2 Mb. A combination of syntenic markers from BTx642 and primary annotated genes from BTx623 were used to identify misjoins in the assembly. Syntenic markers consisted of a total of 32,400 unique, non-repetitive, non-overlapping 1.0 Kb sequences generated using the version 1.0 S. bicolor BTx642 genome release. The gene set consisted of a subset of 12,641 BTx642 annotated primary genes from the V1 annotation. Both the syntenic markers and the primary gene set were aligned to the polished assembly and a total of 5 misjoins were identified in the assembly. Contigs were then ordered, oriented, and assembled into chromosomes using the combination of syntenic markers and genes. Contigs terminating in significant telomeric sequence were identified using the (TTTAGGG)n repeat, and care was taken to make sure that they were properly oriented in the release assembly. A total of 105 joins were applied to the assembly to form the final assembly consisting of 10 chromosomes. Each chromosome join is padded with 10,000 Ns. Chromosomes were numbered using the V1.0 S. bicolor BTx642 assembly available from Phytozome (Goodstein et al., 2012). The remaining scaffolds were screened against bacterial proteins, organelle sequences, GenBank nr and removed if found to be a contaminant. Chloroplast and mitochondrial genomes were generated using the OatK pipeline (Zhou et al., 2024). Homozygous SNPs and INDELs were corrected in the release using 44.5x of the Illumina fragment reads (2x150, 400bp insert) by aligning the reads using bwa mem (Li, 2013) and identifying homozygous SNPs and INDELs with the GATK's UnifiedGenotyper tool (McKenna et al., 2010). A total of 9,558 homozygous snps and 5,812 homozygous indels were corrected in the consensus sequence. The final release consisted of 693.1 Mb of sequence with a contig N50 of 18.7 Mb.

PI_569459 (LUWAIAH / IS 19523) | The PI_569459 (LUWAIAH / IS 19523) assembly was generated by assembling the PACBIO CLR (55.74x coverage, average read size 8,206 bp) reads using the CANU (Koren et al., 2017) assembler and subsequently polished using ARROW (Chin et al., 2013). This produced an initial assembly consisting of 778.3 Mb sequence assembled and a contig N50 of 2.9 Mb. A combination of syntenic markers from BTx642 and primary annotated genes from BTx623 were used to identify misjoins in the assembly. Syntenic markers consisted of a total of 32,400 unique, non-repetitive, non-overlapping 1.0 Kb sequences generated using the version 1.0 S. bicolor BTx642 genome release. The gene set consisted of a subset of 12,641 BTx642 annotated primary genes from the V1 annotation. Both the syntenic markers and the primary gene set were aligned to the polished assembly and a total of 3 misjoins were identified in the assembly. Contigs were then ordered, oriented, and assembled into chromosomes using the combination of syntenic markers and genes. Contigs terminating in significant telomeric sequence were identified using the (TTTAGGG)n repeat, and care was taken to make sure that they were properly oriented in the release assembly. A total of 453 joins were applied to the assembly to form the final assembly consisting of 10 chromosomes. Each chromosome join is padded with 10,000 Ns. Chromosomes were numbered using the V1.0 S. bicolor BTx642 assembly available from Phytozome (Goodstein et al., 2012). The remaining scaffolds were screened against bacterial proteins, organelle sequences, GenBank nr and removed if found to be a contaminant. Chloroplast and mitochondrial genomes were generated using the OatK pipeline (Zhou et al., 2024). Homozygous SNPs and INDELs were corrected in the release using 37.1x of the Illumina fragment reads (2x150, 400bp insert) by aligning the reads using bwa mem (Li, 2013) and identifying homozygous SNPs and INDELs with the GATK's UnifiedGenotyper tool (McKenna et al., 2010). A total of 2,612 homozygous snps and 19,412 homozygous indels were corrected in the consensus sequence. The final release consisted of 712.9 Mb of sequence with a contig N50 of 5.3 Mb.

PI_570071 (IS 22849) | The PI_570071 (IS 22849) assembly was generated by assembling the PACBIO CLR (48.39x coverage, average read size 7,673 bp) reads using the CANU (Koren et al., 2017) assembler and subsequently polished using ARROW (Chin et al., 2013). This produced an initial assembly consisting of 715.0 Mb sequence assembled and a contig N50 of 3.0 Mb. A combination of syntenic markers from BTx642 and primary annotated genes from BTx623 were used to identify misjoins in the assembly. Syntenic markers consisted of a total of 32,400 unique, non-repetitive, non-overlapping 1.0 Kb sequences generated using the version 1.0 S. bicolor BTx642 genome release. The gene set consisted of a subset of 12,641 BTx642 annotated primary genes from the V1 annotation. Both the syntenic markers and the primary gene set were aligned to the polished assembly and a total of 4 misjoins were identified in the assembly. Contigs were then ordered, oriented, and assembled into chromosomes using the combination of syntenic markers and genes. Contigs terminating in significant telomeric sequence were identified using the (TTTAGGG)n repeat, and care was taken to make sure that they were properly oriented in the release assembly. A total of 563 joins were applied to the assembly to form the final assembly consisting of 10 chromosomes. Each chromosome join is padded with 10,000 Ns. Chromosomes were numbered using the V1.0 S. bicolor BTx642 assembly available from Phytozome (Goodstein et al., 2012). The remaining scaffolds were screened against bacterial proteins, organelle sequences, GenBank nr and removed if found to be a contaminant. Chloroplast and mitochondrial genomes were generated using the OatK pipeline (Zhou et al., 2024). Homozygous SNPs and INDELs were corrected in the release using 51.7x of the Illumina fragment reads (2x150, 400bp insert) by aligning the reads using bwa mem (Li, 2013) and identifying homozygous SNPs and INDELs with the GATK's UnifiedGenotyper tool (McKenna et al., 2010). A total of 212 homozygous snps and 5,761 homozygous indels were corrected in the consensus sequence. The final release consisted of 667.5 Mb of sequence with a contig N50 of 4.3 Mb.

PI_656031 (IRAT204) | The PI_656031 (IRAT204) assembly was generated by assembling the PACBIO CLR (51.19x coverage, average read size 14,950 bp) reads using the CANU (Koren et al., 2017) assembler and subsequently polished using ARROW (Chin et al., 2013). This produced an initial assembly consisting of 826.0 Mb sequence assembled and a contig N50 of 8.1 Mb. A combination of syntenic markers from BTx642 and primary annotated genes from BTx623 were used to identify misjoins in the assembly. Syntenic markers consisted of a total of 32,400 unique, non-repetitive, non-overlapping 1.0 Kb sequences generated using the version 1.0 S. bicolor BTx642 genome release. The gene set consisted of a subset of 12,641 BTx642 annotated primary genes from the V1 annotation. Both the syntenic markers and the primary gene set were aligned to the polished assembly and a total of 2 misjoins were identified in the assembly. Contigs were then ordered, oriented, and assembled into chromosomes using the combination of syntenic markers and genes. Contigs terminating in significant telomeric sequence were identified using the (TTTAGGG)n repeat, and care was taken to make sure that they were properly oriented in the release assembly. A total of 253 joins were applied to the assembly to form the final assembly consisting of 10 chromosomes. Each chromosome join is padded with 10,000 Ns. Chromosomes were numbered using the V1.0 S. bicolor BTx642 assembly available from Phytozome (Goodstein et al., 2012). The remaining scaffolds were screened against bacterial proteins, organelle sequences, GenBank nr and removed if found to be a contaminant. Chloroplast and mitochondrial genomes were generated using the OatK pipeline (Zhou et al., 2024). Homozygous SNPs and INDELs were corrected in the release using 118.4x of the Illumina fragment reads (2x150, 400bp insert) by aligning the reads using bwa mem (Li, 2013) and identifying homozygous SNPs and INDELs with the GATK's UnifiedGenotyper tool (McKenna et al., 2010). A total of 243 homozygous snps and 22,710 homozygous indels were corrected in the consensus sequence. The final release consisted of 741.4 Mb of sequence with a contig N50 of 20.7 Mb.

PI_656044 (Kuyuma) | The PI_656044 (Kuyuma) assembly was generated by assembling the PACBIO CLR (44.34x coverage, average read size 9,056 bp) reads using the CANU (Koren et al., 2017) assembler and subsequently polished using ARROW (Chin et al., 2013). This produced an initial assembly consisting of 727.0 Mb sequence assembled and a contig N50 of 7.1 Mb. A combination of syntenic markers from BTx642 and primary annotated genes from BTx623 were used to identify misjoins in the assembly. Syntenic markers consisted of a total of 32,400 unique, non-repetitive, non-overlapping 1.0 Kb sequences generated using the version 1.0 S. bicolor BTx642 genome release. The gene set consisted of a subset of 12,641 BTx642 annotated primary genes from the V1 annotation. Both the syntenic markers and the primary gene set were aligned to the polished assembly and a total of 3 misjoins were identified in the assembly. Contigs were then ordered, oriented, and assembled into chromosomes using the combination of syntenic markers and genes. Contigs terminating in significant telomeric sequence were identified using the (TTTAGGG)n repeat, and care was taken to make sure that they were properly oriented in the release assembly. A total of 186 joins were applied to the assembly to form the final assembly consisting of 10 chromosomes. Each chromosome join is padded with 10,000 Ns. Chromosomes were numbered using the V1.0 S. bicolor BTx642 assembly available from Phytozome (Goodstein et al., 2012). The remaining scaffolds were screened against bacterial proteins, organelle sequences, GenBank nr and removed if found to be a contaminant. Chloroplast and mitochondrial genomes were generated using the OatK pipeline (Zhou et al., 2024). Homozygous SNPs and INDELs were corrected in the release using 43.3x of the Illumina fragment reads (2x150, 400bp insert) by aligning the reads using bwa mem (Li, 2013) and identifying homozygous SNPs and INDELs with the GATK's UnifiedGenotyper tool (McKenna et al., 2010). A total of 157 homozygous snps and 8,052 homozygous indels were corrected in the consensus sequence. The final release consisted of 685.4 Mb of sequence with a contig N50 of 14.1 Mb.

PI_154844 (GRASSL) | The PI_154844 (GRASSL) assembly was generated by assembling the PACBIO CLR (49.40x coverage, average read size 14,439 bp) reads using the CANU (Koren et al., 2017) assembler and subsequently polished using ARROW (Chin et al., 2013). This produced an initial assembly consisting of 741.5 Mb sequence assembled and a contig N50 of 13.6 Mb. A combination of syntenic markers from BTx642 and primary annotated genes from BTx623 were used to identify misjoins in the assembly. Syntenic markers consisted of a total of 32,400 unique, non-repetitive, non-overlapping 1.0 Kb sequences generated using the version 1.0 S. bicolor BTx642 genome release. The gene set consisted of a subset of 12,641 BTx642 annotated primary genes from the V1 annotation. Both the syntenic markers and the primary gene set were aligned to the polished assembly and a total of 2 misjoins were identified in the assembly. Contigs were then ordered, oriented, and assembled into chromosomes using the combination of syntenic markers and genes. Contigs terminating in significant telomeric sequence were identified using the (TTTAGGG)n repeat, and care was taken to make sure that they were properly oriented in the release assembly. A total of 87 joins were applied to the assembly to form the final assembly consisting of 10 chromosomes. Each chromosome join is padded with 10,000 Ns. Chromosomes were numbered using the V1.0 S. bicolor BTx642 assembly available from Phytozome (Goodstein et al., 2012). The remaining scaffolds were screened against bacterial proteins, organelle sequences, GenBank nr and removed if found to be a contaminant. Chloroplast and mitochondrial genomes were generated using the OatK pipeline (Zhou et al., 2024). Homozygous SNPs and INDELs were corrected in the release using 37.0x of the Illumina fragment reads (2x150, 400bp insert) by aligning the reads using bwa mem (Li, 2013) and identifying homozygous SNPs and INDELs with the GATK's UnifiedGenotyper tool (McKenna et al., 2010). A total of 107 homozygous snps and 3,644 homozygous indels were corrected in the consensus sequence. The final release consisted of 700.8 Mb of sequence with a contig N50 of 28.4 Mb.

PI_276816 (IS 12646) | The PI_276816 (IS 12646) assembly was generated by assembling the PACBIO CLR (77.85x coverage, average read size 13,266 bp) reads using the CANU (Koren et al., 2017) assembler and subsequently polished using ARROW (Chin et al., 2013). This produced an initial assembly consisting of 735.3 Mb sequence assembled and a contig N50 of 13.9 Mb. A combination of syntenic markers from BTx642 and primary annotated genes from BTx623 were used to identify misjoins in the assembly. Syntenic markers consisted of a total of 32,400 unique, non-repetitive, non-overlapping 1.0 Kb sequences generated using the version 1.0 S. bicolor BTx642 genome release. The gene set consisted of a subset of 12,641 BTx642 annotated primary genes from the V1 annotation. Both the syntenic markers and the primary gene set were aligned to the polished assembly and a total of 4 misjoins were identified in the assembly. Contigs were then ordered, oriented, and assembled into chromosomes using the combination of syntenic markers and genes. Contigs terminating in significant telomeric sequence were identified using the (TTTAGGG)n repeat, and care was taken to make sure that they were properly oriented in the release assembly. A total of 97 joins were applied to the assembly to form the final assembly consisting of 10 chromosomes. Each chromosome join is padded with 10,000 Ns. Chromosomes were numbered using the V1.0 S. bicolor BTx642 assembly available from Phytozome (Goodstein et al., 2012). The remaining scaffolds were screened against bacterial proteins, organelle sequences, GenBank nr and removed if found to be a contaminant. Chloroplast and mitochondrial genomes were generated using the OatK pipeline (Zhou et al., 2024). Homozygous SNPs and INDELs were corrected in the release using 51.9x of the Illumina fragment reads (2x150, 400bp insert) by aligning the reads using bwa mem (Li, 2013) and identifying homozygous SNPs and INDELs with the GATK's UnifiedGenotyper tool (McKenna et al., 2010). A total of 59 homozygous snps and 6,164 homozygous indels were corrected in the consensus sequence. The final release consisted of 698.7 Mb of sequence with a contig N50 of 21.4 Mb.

PI_329501 (IS 11257) | The PI_329501 (IS 11257) assembly was generated by assembling the PACBIO CLR (64.83x coverage, average read size 8,236 bp) reads using the CANU (Koren et al., 2017) assembler and subsequently polished using ARROW (Chin et al., 2013). This produced an initial assembly consisting of 960.5 Mb sequence assembled and a contig N50 of 1.1 Mb. A combination of syntenic markers from BTx642 and primary annotated genes from BTx623 were used to identify misjoins in the assembly. Syntenic markers consisted of a total of 32,400 unique, non-repetitive, non-overlapping 1.0 Kb sequences generated using the version 1.0 S. bicolor BTx642 genome release. The gene set consisted of a subset of 12,641 BTx642 annotated primary genes from the V1 annotation. Both the syntenic markers and the primary gene set were aligned to the polished assembly and a total of 7 misjoins were identified in the assembly. Contigs were then ordered, oriented, and assembled into chromosomes using the combination of syntenic markers and genes. Contigs terminating in significant telomeric sequence were identified using the (TTTAGGG)n repeat, and care was taken to make sure that they were properly oriented in the release assembly. A total of 789 joins were applied to the assembly to form the final assembly consisting of 10 chromosomes. Each chromosome join is padded with 10,000 Ns. Chromosomes were numbered using the V1.0 S. bicolor BTx642 assembly available from Phytozome (Goodstein et al., 2012). The remaining scaffolds were screened against bacterial proteins, organelle sequences, GenBank nr and removed if found to be a contaminant. Chloroplast and mitochondrial genomes were generated using the OatK pipeline (Zhou et al., 2024). Homozygous SNPs and INDELs were corrected in the release using 41.5x of the Illumina fragment reads (2x150, 400bp insert) by aligning the reads using bwa mem (Li, 2013) and identifying homozygous SNPs and INDELs with the GATK's UnifiedGenotyper tool (McKenna et al., 2010). A total of 108 homozygous snps and 6,821 homozygous indels were corrected in the consensus sequence. The final release consisted of 760.5 Mb of sequence with a contig N50 of 2.6 Mb.

PI_513676 (Matche-Dakomia / IS 20380) | The PI_513676 (Matche-Dakomia / IS 20380) assembly was generated by assembling the PACBIO CLR (61.76x coverage, average read size 16,595 bp) reads using the CANU (Koren et al., 2017) assembler and subsequently polished using ARROW (Chin et al., 2013). This produced an initial assembly consisting of 732.1 Mb sequence assembled and a contig N50 of 16.6 Mb. A combination of syntenic markers from BTx642 and primary annotated genes from BTx623 were used to identify misjoins in the assembly. Syntenic markers consisted of a total of 32,400 unique, non-repetitive, non-overlapping 1.0 Kb sequences generated using the version 1.0 S. bicolor BTx642 genome release. The gene set consisted of a subset of 12,641 BTx642 annotated primary genes from the V1 annotation. Both the syntenic markers and the primary gene set were aligned to the polished assembly and a total of no misjoins were identified in the assembly. Contigs were then ordered, oriented, and assembled into chromosomes using the combination of syntenic markers and genes. Contigs terminating in significant telomeric sequence were identified using the (TTTAGGG)n repeat, and care was taken to make sure that they were properly oriented in the release assembly. A total of 63 joins were applied to the assembly to form the final assembly consisting of 10 chromosomes. Each chromosome join is padded with 10,000 Ns. Chromosomes were numbered using the V1.0 S. bicolor BTx642 assembly available from Phytozome (Goodstein et al., 2012). The remaining scaffolds were screened against bacterial proteins, organelle sequences, GenBank nr and removed if found to be a contaminant. Chloroplast and mitochondrial genomes were generated using the OatK pipeline (Zhou et al., 2024). Homozygous SNPs and INDELs were corrected in the release using 49.6x of the Illumina fragment reads (2x150, 400bp insert) by aligning the reads using bwa mem (Li, 2013) and identifying homozygous SNPs and INDELs with the GATK's UnifiedGenotyper tool (McKenna et al., 2010). A total of 95 homozygous snps and 3,100 homozygous indels were corrected in the consensus sequence. The final release consisted of 688.5 Mb of sequence with a contig N50 of 29.3 Mb.

PI_533766 (SC 265) | The PI_533766 (SC 265) assembly was generated by assembling the PACBIO CLR (82.82x coverage, average read size 14,269 bp) reads using the CANU (Koren et al., 2017) assembler and subsequently polished using ARROW (Chin et al., 2013). This produced an initial assembly consisting of 727.2 Mb sequence assembled and a contig N50 of 11.1 Mb. A combination of syntenic markers from BTx642 and primary annotated genes from BTx623 were used to identify misjoins in the assembly. Syntenic markers consisted of a total of 32,400 unique, non-repetitive, non-overlapping 1.0 Kb sequences generated using the version 1.0 S. bicolor BTx642 genome release. The gene set consisted of a subset of 12,641 BTx642 annotated primary genes from the V1 annotation. Both the syntenic markers and the primary gene set were aligned to the polished assembly and a total of 3 misjoins were identified in the assembly. Contigs were then ordered, oriented, and assembled into chromosomes using the combination of syntenic markers and genes. Contigs terminating in significant telomeric sequence were identified using the (TTTAGGG)n repeat, and care was taken to make sure that they were properly oriented in the release assembly. A total of 111 joins were applied to the assembly to form the final assembly consisting of 10 chromosomes. Each chromosome join is padded with 10,000 Ns. Chromosomes were numbered using the V1.0 S. bicolor BTx642 assembly available from Phytozome (Goodstein et al., 2012). The remaining scaffolds were screened against bacterial proteins, organelle sequences, GenBank nr and removed if found to be a contaminant. Chloroplast and mitochondrial genomes were generated using the OatK pipeline (Zhou et al., 2024). Homozygous SNPs and INDELs were corrected in the release using 52.0x of the Illumina fragment reads (2x150, 400bp insert) by aligning the reads using bwa mem (Li, 2013) and identifying homozygous SNPs and INDELs with the GATK's UnifiedGenotyper tool (McKenna et al., 2010). A total of 66 homozygous snps and 5,701 homozygous indels were corrected in the consensus sequence. The final release consisted of 687.2 Mb of sequence with a contig N50 of 22.4 Mb.

PI_534133 (SC 35) | The PI_534133 (SC 35) assembly was generated by assembling the PACBIO CLR (70.32x coverage, average read size 13,048 bp) reads using the CANU (Koren et al., 2017) assembler and subsequently polished using ARROW (Chin et al., 2013). This produced an initial assembly consisting of 726.3 Mb sequence assembled and a contig N50 of 5.6 Mb. A combination of syntenic markers from BTx642 and primary annotated genes from BTx623 were used to identify misjoins in the assembly. Syntenic markers consisted of a total of 32,400 unique, non-repetitive, non-overlapping 1.0 Kb sequences generated using the version 1.0 S. bicolor BTx642 genome release. The gene set consisted of a subset of 12,641 BTx642 annotated primary genes from the V1 annotation. Both the syntenic markers and the primary gene set were aligned to the polished assembly and a total of 1 misjoins were identified in the assembly. Contigs were then ordered, oriented, and assembled into chromosomes using the combination of syntenic markers and genes. Contigs terminating in significant telomeric sequence were identified using the (TTTAGGG)n repeat, and care was taken to make sure that they were properly oriented in the release assembly. A total of 206 joins were applied to the assembly to form the final assembly consisting of 10 chromosomes. Each chromosome join is padded with 10,000 Ns. Chromosomes were numbered using the V1.0 S. bicolor BTx642 assembly available from Phytozome (Goodstein et al., 2012). The remaining scaffolds were screened against bacterial proteins, organelle sequences, GenBank nr and removed if found to be a contaminant. Chloroplast and mitochondrial genomes were generated using the OatK pipeline (Zhou et al., 2024). Homozygous SNPs and INDELs were corrected in the release using 52.4x of the Illumina fragment reads (2x150, 400bp insert) by aligning the reads using bwa mem (Li, 2013) and identifying homozygous SNPs and INDELs with the GATK's UnifiedGenotyper tool (McKenna et al., 2010). A total of 174 homozygous snps and 3,055 homozygous indels were corrected in the consensus sequence. The final release consisted of 687.5 Mb of sequence with a contig N50 of 15.4 Mb.

PI_565121 (MACIA) | The PI_565121 (MACIA) assembly was generated by assembling the PACBIO CLR (51.46x coverage, average read size 14,786 bp) reads using the MECAT (Xiao et al., 2017) assembler and subsequently polished using ARROW (Chin et al., 2013). This produced an initial assembly consisting of 715.8 Mb sequence assembled and a contig N50 of 25.0 Mb. A combination of syntenic markers from BTx642 and primary annotated genes from BTx623 were used to identify misjoins in the assembly. Syntenic markers consisted of a total of 32,400 unique, non-repetitive, non-overlapping 1.0 Kb sequences generated using the version 1.0 S. bicolor BTx642 genome release. The gene set consisted of a subset of 12,641 BTx642 annotated primary genes from the V1 annotation. Both the syntenic markers and the primary gene set were aligned to the polished assembly and a total of 2 misjoins were identified in the assembly. Contigs were then ordered, oriented, and assembled into chromosomes using the combination of syntenic markers and genes. Contigs terminating in significant telomeric sequence were identified using the (TTTAGGG)n repeat, and care was taken to make sure that they were properly oriented in the release assembly. A total of 85 joins were applied to the assembly to form the final assembly consisting of 10 chromosomes. Each chromosome join is padded with 10,000 Ns. Chromosomes were numbered using the V1.0 S. bicolor BTx642 assembly available from Phytozome (Goodstein et al., 2012). The remaining scaffolds were screened against bacterial proteins, organelle sequences, GenBank nr and removed if found to be a contaminant. Chloroplast and mitochondrial genomes were generated using the OatK pipeline (Zhou et al., 2024). Homozygous SNPs and INDELs were corrected in the release using 45.6x of the Illumina fragment reads (2x150, 400bp insert) by aligning the reads using bwa mem (Li, 2013) and identifying homozygous SNPs and INDELs with the GATK's UnifiedGenotyper tool (McKenna et al., 2010). A total of 675 homozygous snps and 4,838 homozygous indels were corrected in the consensus sequence. The final release consisted of 688.4 Mb of sequence with a contig N50 of 26.1 Mb.

PI_576434 (SC 1103) | The PI_576434 (SC 1103) assembly was generated by assembling the PACBIO CLR (63.40x coverage, average read size 10,082 bp) reads using the CANU (Koren et al., 2017) assembler and subsequently polished using ARROW (Chin et al., 2013). This produced an initial assembly consisting of 733.7 Mb sequence assembled and a contig N50 of 8.0 Mb. A combination of syntenic markers from BTx642 and primary annotated genes from BTx623 were used to identify misjoins in the assembly. Syntenic markers consisted of a total of 32,400 unique, non-repetitive, non-overlapping 1.0 Kb sequences generated using the version 1.0 S. bicolor BTx642 genome release. The gene set consisted of a subset of 12,641 BTx642 annotated primary genes from the V1 annotation. Both the syntenic markers and the primary gene set were aligned to the polished assembly and a total of no misjoins were identified in the assembly. Contigs were then ordered, oriented, and assembled into chromosomes using the combination of syntenic markers and genes. Contigs terminating in significant telomeric sequence were identified using the (TTTAGGG)n repeat, and care was taken to make sure that they were properly oriented in the release assembly. A total of 162 joins were applied to the assembly to form the final assembly consisting of 10 chromosomes. Each chromosome join is padded with 10,000 Ns. Chromosomes were numbered using the V1.0 S. bicolor BTx642 assembly available from Phytozome (Goodstein et al., 2012). The remaining scaffolds were screened against bacterial proteins, organelle sequences, GenBank nr and removed if found to be a contaminant. Chloroplast and mitochondrial genomes were generated using the OatK pipeline (Zhou et al., 2024). Homozygous SNPs and INDELs were corrected in the release using 49.9x of the Illumina fragment reads (2x150, 400bp insert) by aligning the reads using bwa mem (Li, 2013) and identifying homozygous SNPs and INDELs with the GATK's UnifiedGenotyper tool (McKenna et al., 2010). A total of 94 homozygous snps and 3,503 homozygous indels were corrected in the consensus sequence. The final release consisted of 687.8 Mb of sequence with a contig N50 of 20.1 Mb.

PI_585966 (DJOFELA/ IS 26395) | The PI_585966 (DJOFELA/ IS 26395) assembly was generated by assembling the PACBIO CLR (53.59x coverage, average read size 9,149 bp) reads using the CANU (Koren et al., 2017) assembler and subsequently polished using ARROW (Chin et al., 2013). This produced an initial assembly consisting of 775.6 Mb sequence assembled and a contig N50 of 4.6 Mb. A combination of syntenic markers from BTx642 and primary annotated genes from BTx623 were used to identify misjoins in the assembly. Syntenic markers consisted of a total of 32,400 unique, non-repetitive, non-overlapping 1.0 Kb sequences generated using the version 1.0 S. bicolor BTx642 genome release. The gene set consisted of a subset of 12,641 BTx642 annotated primary genes from the V1 annotation. Both the syntenic markers and the primary gene set were aligned to the polished assembly and a total of 12 misjoins were identified in the assembly. Contigs were then ordered, oriented, and assembled into chromosomes using the combination of syntenic markers and genes. Contigs terminating in significant telomeric sequence were identified using the (TTTAGGG)n repeat, and care was taken to make sure that they were properly oriented in the release assembly. A total of 246 joins were applied to the assembly to form the final assembly consisting of 10 chromosomes. Each chromosome join is padded with 10,000 Ns. Chromosomes were numbered using the V1.0 S. bicolor BTx642 assembly available from Phytozome (Goodstein et al., 2012). The remaining scaffolds were screened against bacterial proteins, organelle sequences, GenBank nr and removed if found to be a contaminant. Chloroplast and mitochondrial genomes were generated using the OatK pipeline (Zhou et al., 2024). Homozygous SNPs and INDELs were corrected in the release using 44.8x of the Illumina fragment reads (2x150, 400bp insert) by aligning the reads using bwa mem (Li, 2013) and identifying homozygous SNPs and INDELs with the GATK's UnifiedGenotyper tool (McKenna et al., 2010). A total of 45,879 homozygous snps and 11,557 homozygous indels were corrected in the consensus sequence. The final release consisted of 679.6 Mb of sequence with a contig N50 of 9.5 Mb.

PI_597980 (SC 1345/CSM-90) | The PI_597980 (SC 1345/CSM-90) assembly was generated by assembling the PACBIO CLR (63.87x coverage, average read size 10,732 bp) reads using the CANU (Koren et al., 2017) assembler and subsequently polished using ARROW (Chin et al., 2013). This produced an initial assembly consisting of 737.6 Mb sequence assembled and a contig N50 of 13.1 Mb. A combination of syntenic markers from BTx642 and primary annotated genes from BTx623 were used to identify misjoins in the assembly. Syntenic markers consisted of a total of 32,400 unique, non-repetitive, non-overlapping 1.0 Kb sequences generated using the version 1.0 S. bicolor BTx642 genome release. The gene set consisted of a subset of 12,641 BTx642 annotated primary genes from the V1 annotation. Both the syntenic markers and the primary gene set were aligned to the polished assembly and a total of 2 misjoins were identified in the assembly. Contigs were then ordered, oriented, and assembled into chromosomes using the combination of syntenic markers and genes. Contigs terminating in significant telomeric sequence were identified using the (TTTAGGG)n repeat, and care was taken to make sure that they were properly oriented in the release assembly. A total of 101 joins were applied to the assembly to form the final assembly consisting of 10 chromosomes. Each chromosome join is padded with 10,000 Ns. Chromosomes were numbered using the V1.0 S. bicolor BTx642 assembly available from Phytozome (Goodstein et al., 2012). The remaining scaffolds were screened against bacterial proteins, organelle sequences, GenBank nr and removed if found to be a contaminant. Chloroplast and mitochondrial genomes were generated using the OatK pipeline (Zhou et al., 2024). Homozygous SNPs and INDELs were corrected in the release using 49.2x of the Illumina fragment reads (2x150, 400bp insert) by aligning the reads using bwa mem (Li, 2013) and identifying homozygous SNPs and INDELs with the GATK's UnifiedGenotyper tool (McKenna et al., 2010). A total of 25 homozygous snps and 3,058 homozygous indels were corrected in the consensus sequence. The final release consisted of 691.6 Mb of sequence with a contig N50 of 23.0 Mb.

PI_655981 (CSM-63, SAP-50) | The PI_655981 (CSM-63) assembly was generated by assembling the PACBIO CLR (52.37x coverage, average read size 14,721 bp) reads using the CANU (Koren et al., 2017) assembler and subsequently polished using ARROW (Chin et al., 2013). This produced an initial assembly consisting of 733.0 Mb sequence assembled and a contig N50 of 15.3 Mb. A combination of syntenic markers from BTx642 and primary annotated genes from BTx623 were used to identify misjoins in the assembly. Syntenic markers consisted of a total of 32,400 unique, non-repetitive, non-overlapping 1.0 Kb sequences generated using the version 1.0 S. bicolor BTx642 genome release. The gene set consisted of a subset of 12,641 BTx642 annotated primary genes from the V1 annotation. Both the syntenic markers and the primary gene set were aligned to the polished assembly and a total of 2 misjoins were identified in the assembly. Contigs were then ordered, oriented, and assembled into chromosomes using the combination of syntenic markers and genes. Contigs terminating in significant telomeric sequence were identified using the (TTTAGGG)n repeat, and care was taken to make sure that they were properly oriented in the release assembly. A total of 85 joins were applied to the assembly to form the final assembly consisting of 10 chromosomes. Each chromosome join is padded with 10,000 Ns. Chromosomes were numbered using the V1.0 S. bicolor BTx642 assembly available from Phytozome (Goodstein et al., 2012). The remaining scaffolds were screened against bacterial proteins, organelle sequences, GenBank nr and removed if found to be a contaminant. Chloroplast and mitochondrial genomes were generated using the OatK pipeline (Zhou et al., 2024). Homozygous SNPs and INDELs were corrected in the release using 11.9x of the Illumina fragment reads (2x150, 400bp insert) by aligning the reads using bwa mem (Li, 2013) and identifying homozygous SNPs and INDELs with the GATK's UnifiedGenotyper tool (McKenna et al., 2010). A total of 14,986 homozygous snps and 6,110 homozygous indels were corrected in the consensus sequence. The final release consisted of 691.5 Mb of sequence with a contig N50 of 22.0 Mb.

PI_655988 (COMBINE KAFIR-60) | The PI_655988 (COMBINE KAFIR-60) assembly was generated by assembling the PACBIO CLR (53.03x coverage, average read size 8,266 bp) reads using the CANU (Koren et al., 2017) assembler and subsequently polished using ARROW (Chin et al., 2013). This produced an initial assembly consisting of 733.6 Mb sequence assembled and a contig N50 of 12.9 Mb. A combination of syntenic markers from BTx642 and primary annotated genes from BTx623 were used to identify misjoins in the assembly. Syntenic markers consisted of a total of 32,400 unique, non-repetitive, non-overlapping 1.0 Kb sequences generated using the version 1.0 S. bicolor BTx642 genome release. The gene set consisted of a subset of 12,641 BTx642 annotated primary genes from the V1 annotation. Both the syntenic markers and the primary gene set were aligned to the polished assembly and a total of no misjoins were identified in the assembly. Contigs were then ordered, oriented, and assembled into chromosomes using the combination of syntenic markers and genes. Contigs terminating in significant telomeric sequence were identified using the (TTTAGGG)n repeat, and care was taken to make sure that they were properly oriented in the release assembly. A total of 90 joins were applied to the assembly to form the final assembly consisting of 10 chromosomes. Each chromosome join is padded with 10,000 Ns. Chromosomes were numbered using the V1.0 S. bicolor BTx642 assembly available from Phytozome (Goodstein et al., 2012). The remaining scaffolds were screened against bacterial proteins, organelle sequences, GenBank nr and removed if found to be a contaminant. Chloroplast and mitochondrial genomes were generated using the OatK pipeline (Zhou et al., 2024). Homozygous SNPs and INDELs were corrected in the release using 42.5x of the Illumina fragment reads (2x150, 400bp insert) by aligning the reads using bwa mem (Li, 2013) and identifying homozygous SNPs and INDELs with the GATK's UnifiedGenotyper tool (McKenna et al., 2010). A total of 54 homozygous snps and 356 homozygous indels were corrected in the consensus sequence. The final release consisted of 696.3 Mb of sequence with a contig N50 of 24.3 Mb.

PI_656015 (Ajabsido) | The PI_656015 (Ajabsido) assembly was generated by assembling the PACBIO CLR (55.71x coverage, average read size 9,724 bp) reads using the CANU (Koren et al., 2017) assembler and subsequently polished using ARROW (Chin et al., 2013). This produced an initial assembly consisting of 733.6 Mb sequence assembled and a contig N50 of 13.2 Mb. A combination of syntenic markers from BTx642 and primary annotated genes from BTx623 were used to identify misjoins in the assembly. Syntenic markers consisted of a total of 32,400 unique, non-repetitive, non-overlapping 1.0 Kb sequences generated using the version 1.0 S. bicolor BTx642 genome release. The gene set consisted of a subset of 12,641 BTx642 annotated primary genes from the V1 annotation. Both the syntenic markers and the primary gene set were aligned to the polished assembly and a total of 2 misjoins were identified in the assembly. Contigs were then ordered, oriented, and assembled into chromosomes using the combination of syntenic markers and genes. Contigs terminating in significant telomeric sequence were identified using the (TTTAGGG)n repeat, and care was taken to make sure that they were properly oriented in the release assembly. A total of 96 joins were applied to the assembly to form the final assembly consisting of 10 chromosomes. Each chromosome join is padded with 10,000 Ns. Chromosomes were numbered using the V1.0 S. bicolor BTx642 assembly available from Phytozome (Goodstein et al., 2012). The remaining scaffolds were screened against bacterial proteins, organelle sequences, GenBank nr and removed if found to be a contaminant. Chloroplast and mitochondrial genomes were generated using the OatK pipeline (Zhou et al., 2024). Homozygous SNPs and INDELs were corrected in the release using 19.6x of the Illumina fragment reads (2x150, 400bp insert) by aligning the reads using bwa mem (Li, 2013) and identifying homozygous SNPs and INDELs with the GATK's UnifiedGenotyper tool (McKenna et al., 2010). A total of 153 homozygous snps and 1,421 homozygous indels were corrected in the consensus sequence. The final release consisted of 688.8 Mb of sequence with a contig N50 of 22.6 Mb.

PI_656023 (Segaolane) | The PI_656023 (Segaolane) assembly was generated by assembling the PACBIO CLR (88.11x coverage, average read size 15,195 bp) reads using the CANU (Koren et al., 2017) assembler and subsequently polished using ARROW (Chin et al., 2013). This produced an initial assembly consisting of 729.1 Mb sequence assembled and a contig N50 of 14.0 Mb. A combination of syntenic markers from BTx642 and primary annotated genes from BTx623 were used to identify misjoins in the assembly. Syntenic markers consisted of a total of 32,400 unique, non-repetitive, non-overlapping 1.0 Kb sequences generated using the version 1.0 S. bicolor BTx642 genome release. The gene set consisted of a subset of 12,641 BTx642 annotated primary genes from the V1 annotation. Both the syntenic markers and the primary gene set were aligned to the polished assembly and a total of 5 misjoins were identified in the assembly. Contigs were then ordered, oriented, and assembled into chromosomes using the combination of syntenic markers and genes. Contigs terminating in significant telomeric sequence were identified using the (TTTAGGG)n repeat, and care was taken to make sure that they were properly oriented in the release assembly. A total of 78 joins were applied to the assembly to form the final assembly consisting of 10 chromosomes. Each chromosome join is padded with 10,000 Ns. Chromosomes were numbered using the V1.0 S. bicolor BTx642 assembly available from Phytozome (Goodstein et al., 2012). The remaining scaffolds were screened against bacterial proteins, organelle sequences, GenBank nr and removed if found to be a contaminant. Chloroplast and mitochondrial genomes were generated using the OatK pipeline (Zhou et al., 2024). Homozygous SNPs and INDELs were corrected in the release using 52.0x of the Illumina fragment reads (2x150, 400bp insert) by aligning the reads using bwa mem (Li, 2013) and identifying homozygous SNPs and INDELs with the GATK's UnifiedGenotyper tool (McKenna et al., 2010). A total of 28 homozygous snps and 6,057 homozygous indels were corrected in the consensus sequence. The final release consisted of 695.7 Mb of sequence with a contig N50 of 27.9 Mb.

PI_656027 (SRN39) | The PI_656027 (SRN39) assembly was generated by assembling the PACBIO CLR (62.89x coverage, average read size 12,322 bp) reads using the CANU (Koren et al., 2017) assembler and subsequently polished using ARROW (Chin et al., 2013). This produced an initial assembly consisting of 736.7 Mb sequence assembled and a contig N50 of 9.8 Mb. A combination of syntenic markers from BTx642 and primary annotated genes from BTx623 were used to identify misjoins in the assembly. Syntenic markers consisted of a total of 32,400 unique, non-repetitive, non-overlapping 1.0 Kb sequences generated using the version 1.0 S. bicolor BTx642 genome release. The gene set consisted of a subset of 12,641 BTx642 annotated primary genes from the V1 annotation. Both the syntenic markers and the primary gene set were aligned to the polished assembly and a total of 2 misjoins were identified in the assembly. Contigs were then ordered, oriented, and assembled into chromosomes using the combination of syntenic markers and genes. Contigs terminating in significant telomeric sequence were identified using the (TTTAGGG)n repeat, and care was taken to make sure that they were properly oriented in the release assembly. A total of 102 joins were applied to the assembly to form the final assembly consisting of 10 chromosomes. Each chromosome join is padded with 10,000 Ns. Chromosomes were numbered using the V1.0 S. bicolor BTx642 assembly available from Phytozome (Goodstein et al., 2012). The remaining scaffolds were screened against bacterial proteins, organelle sequences, GenBank nr and removed if found to be a contaminant. Chloroplast and mitochondrial genomes were generated using the OatK pipeline (Zhou et al., 2024). Homozygous SNPs and INDELs were corrected in the release using 49.6x of the Illumina fragment reads (2x150, 400bp insert) by aligning the reads using bwa mem (Li, 2013) and identifying homozygous SNPs and INDELs with the GATK's UnifiedGenotyper tool (McKenna et al., 2010). A total of 39 homozygous snps and 2,772 homozygous indels were corrected in the consensus sequence. The final release consisted of 696.2 Mb of sequence with a contig N50 of 18.3 Mb.

PI_656050 (Mota Maradi) | The PI_656050 (Mota Maradi) assembly was generated by assembling the PACBIO CLR (72.21x coverage, average read size 12,128 bp) reads using the CANU (Koren et al., 2017) assembler and subsequently polished using ARROW (Chin et al., 2013). This produced an initial assembly consisting of 730.4 Mb sequence assembled and a contig N50 of 18.6 Mb. A combination of syntenic markers from BTx642 and primary annotated genes from BTx623 were used to identify misjoins in the assembly. Syntenic markers consisted of a total of 32,400 unique, non-repetitive, non-overlapping 1.0 Kb sequences generated using the version 1.0 S. bicolor BTx642 genome release. The gene set consisted of a subset of 12,641 BTx642 annotated primary genes from the V1 annotation. Both the syntenic markers and the primary gene set were aligned to the polished assembly and a total of 2 misjoins were identified in the assembly. Contigs were then ordered, oriented, and assembled into chromosomes using the combination of syntenic markers and genes. Contigs terminating in significant telomeric sequence were identified using the (TTTAGGG)n repeat, and care was taken to make sure that they were properly oriented in the release assembly. A total of 109 joins were applied to the assembly to form the final assembly consisting of 10 chromosomes. Each chromosome join is padded with 10,000 Ns. Chromosomes were numbered using the V1.0 S. bicolor BTx642 assembly available from Phytozome (Goodstein et al., 2012). The remaining scaffolds were screened against bacterial proteins, organelle sequences, GenBank nr and removed if found to be a contaminant. Chloroplast and mitochondrial genomes were generated using the OatK pipeline (Zhou et al., 2024). Homozygous SNPs and INDELs were corrected in the release using 51.6x of the Illumina fragment reads (2x150, 400bp insert) by aligning the reads using bwa mem (Li, 2013) and identifying homozygous SNPs and INDELs with the GATK's UnifiedGenotyper tool (McKenna et al., 2010). A total of 51 homozygous snps and 3,223 homozygous indels were corrected in the consensus sequence. The final release consisted of 698.1 Mb of sequence with a contig N50 of 20.9 Mb.

PI_656057 (P898012) | The PI_656057 (P898012) assembly was generated by assembling the PACBIO CLR (67.7x coverage, average read size 9,983 bp) reads using the CANU (Koren et al., 2017) assembler and subsequently polished using ARROW (Chin et al., 2013). This produced an initial assembly consisting of 739.5 Mb sequence assembled and a contig N50 of 10.0 Mb. A combination of syntenic markers from BTx642 and primary annotated genes from BTx623 were used to identify misjoins in the assembly. Syntenic markers consisted of a total of 32,400 unique, non-repetitive, non-overlapping 1.0 Kb sequences generated using the version 1.0 S. bicolor BTx642 genome release. The gene set consisted of a subset of 12,641 BTx642 annotated primary genes from the V1 annotation. Both the syntenic markers and the primary gene set were aligned to the polished assembly and a total of 3 misjoins were identified in the assembly. Contigs were then ordered, oriented, and assembled into chromosomes using the combination of syntenic markers and genes. Contigs terminating in significant telomeric sequence were identified using the (TTTAGGG)n repeat, and care was taken to make sure that they were properly oriented in the release assembly. A total of 142 joins were applied to the assembly to form the final assembly consisting of 10 chromosomes. Each chromosome join is padded with 10,000 Ns. Chromosomes were numbered using the V1.0 S. bicolor BTx642 assembly available from Phytozome (Goodstein et al., 2012). The remaining scaffolds were screened against bacterial proteins, organelle sequences, GenBank nr and removed if found to be a contaminant. Chloroplast and mitochondrial genomes were generated using the OatK pipeline (Zhou et al., 2024). Homozygous SNPs and INDELs were corrected in the release using 50.6x of the Illumina fragment reads (2x150, 400bp insert) by aligning the reads using bwa mem (Li, 2013) and identifying homozygous SNPs and INDELs with the GATK's UnifiedGenotyper tool (McKenna et al., 2010). A total of 44 homozygous snps and 2,871 homozygous indels were corrected in the consensus sequence. The final release consisted of 693.9 Mb of sequence with a contig N50 of 17.8 Mb.

PI_656111 (SC 971) | The PI_656111 (SC 971) assembly was generated by assembling the PACBIO CLR (62.89x coverage, average read size 12,489 bp) reads using the CANU (Koren et al., 2017) assembler and subsequently polished using ARROW (Chin et al., 2013). This produced an initial assembly consisting of 742.1 Mb sequence assembled and a contig N50 of 5.5 Mb. A combination of syntenic markers from BTx642 and primary annotated genes from BTx623 were used to identify misjoins in the assembly. Syntenic markers consisted of a total of 32,400 unique, non-repetitive, non-overlapping 1.0 Kb sequences generated using the version 1.0 S. bicolor BTx642 genome release. The gene set consisted of a subset of 12,641 BTx642 annotated primary genes from the V1 annotation. Both the syntenic markers and the primary gene set were aligned to the polished assembly and a total of 4 misjoins were identified in the assembly. Contigs were then ordered, oriented, and assembled into chromosomes using the combination of syntenic markers and genes. Contigs terminating in significant telomeric sequence were identified using the (TTTAGGG)n repeat, and care was taken to make sure that they were properly oriented in the release assembly. A total of 223 joins were applied to the assembly to form the final assembly consisting of 10 chromosomes. Each chromosome join is padded with 10,000 Ns. Chromosomes were numbered using the V1.0 S. bicolor BTx642 assembly available from Phytozome (Goodstein et al., 2012). The remaining scaffolds were screened against bacterial proteins, organelle sequences, GenBank nr and removed if found to be a contaminant. Chloroplast and mitochondrial genomes were generated using the OatK pipeline (Zhou et al., 2024). Homozygous SNPs and INDELs were corrected in the release using 50.2x of the Illumina fragment reads (2x150, 400bp insert) by aligning the reads using bwa mem (Li, 2013) and identifying homozygous SNPs and INDELs with the GATK's UnifiedGenotyper tool (McKenna et al., 2010). A total of 296 homozygous snps and 2,931 homozygous indels were corrected in the consensus sequence. The final release consisted of 689.7 Mb of sequence with a contig N50 of 12.1 Mb.

PI_660557 (IS 6705) | The PI_660557 (IS 6705) assembly was generated by assembling the PACBIO CLR (49.95x coverage, average read size 10,065 bp) reads using the CANU (Koren et al., 2017) assembler and subsequently polished using ARROW (Chin et al., 2013). This produced an initial assembly consisting of 735.4 Mb sequence assembled and a contig N50 of 12.1 Mb. A combination of syntenic markers from BTx642 and primary annotated genes from BTx623 were used to identify misjoins in the assembly. Syntenic markers consisted of a total of 32,400 unique, non-repetitive, non-overlapping 1.0 Kb sequences generated using the version 1.0 S. bicolor BTx642 genome release. The gene set consisted of a subset of 12,641 BTx642 annotated primary genes from the V1 annotation. Both the syntenic markers and the primary gene set were aligned to the polished assembly and a total of 3 misjoins were identified in the assembly. Contigs were then ordered, oriented, and assembled into chromosomes using the combination of syntenic markers and genes. Contigs terminating in significant telomeric sequence were identified using the (TTTAGGG)n repeat, and care was taken to make sure that they were properly oriented in the release assembly. A total of 97 joins were applied to the assembly to form the final assembly consisting of 10 chromosomes. Each chromosome join is padded with 10,000 Ns. Chromosomes were numbered using the V1.0 S. bicolor BTx642 assembly available from Phytozome (Goodstein et al., 2012). The remaining scaffolds were screened against bacterial proteins, organelle sequences, GenBank nr and removed if found to be a contaminant. Chloroplast and mitochondrial genomes were generated using the OatK pipeline (Zhou et al., 2024). Homozygous SNPs and INDELs were corrected in the release using 47.5x of the Illumina fragment reads (2x150, 400bp insert) by aligning the reads using bwa mem (Li, 2013) and identifying homozygous SNPs and INDELs with the GATK's UnifiedGenotyper tool (McKenna et al., 2010). A total of 85 homozygous snps and 3,665 homozygous indels were corrected in the consensus sequence. The final release consisted of 689.6 Mb of sequence with a contig N50 of 21.5 Mb.

PI_660563 (IS 7173) | The PI_660563 (IS 7173) assembly was generated by assembling the PACBIO CLR (81.82x coverage, average read size 14,797 bp) reads using the CANU (Koren et al., 2017) assembler and subsequently polished using ARROW (Chin et al., 2013). This produced an initial assembly consisting of 737.1 Mb sequence assembled and a contig N50 of 12.1 Mb. A combination of syntenic markers from BTx642 and primary annotated genes from BTx623 were used to identify misjoins in the assembly. Syntenic markers consisted of a total of 32,400 unique, non-repetitive, non-overlapping 1.0 Kb sequences generated using the version 1.0 S. bicolor BTx642 genome release. The gene set consisted of a subset of 12,641 BTx642 annotated primary genes from the V1 annotation. Both the syntenic markers and the primary gene set were aligned to the polished assembly and a total of 2 misjoins were identified in the assembly. Contigs were then ordered, oriented, and assembled into chromosomes using the combination of syntenic markers and genes. Contigs terminating in significant telomeric sequence were identified using the (TTTAGGG)n repeat, and care was taken to make sure that they were properly oriented in the release assembly. A total of 94 joins were applied to the assembly to form the final assembly consisting of 10 chromosomes. Each chromosome join is padded with 10,000 Ns. Chromosomes were numbered using the V1.0 S. bicolor BTx642 assembly available from Phytozome (Goodstein et al., 2012). The remaining scaffolds were screened against bacterial proteins, organelle sequences, GenBank nr and removed if found to be a contaminant. Chloroplast and mitochondrial genomes were generated using the OatK pipeline (Zhou et al., 2024). Homozygous SNPs and INDELs were corrected in the release using 51.3x of the Illumina fragment reads (2x150, 400bp insert) by aligning the reads using bwa mem (Li, 2013) and identifying homozygous SNPs and INDELs with the GATK's UnifiedGenotyper tool (McKenna et al., 2010). A total of 69 homozygous snps and 2,812 homozygous indels were corrected in the consensus sequence. The final release consisted of 697.7 Mb of sequence with a contig N50 of 27.7 Mb.

PI_660565 (Kamjin / IS 3620) | The PI_660565 (Kamjin / IS 3620) assembly was generated by assembling the PACBIO CLR (87.76x coverage, average read size 16,316 bp) reads using the CANU (Koren et al., 2017) assembler and subsequently polished using ARROW (Chin et al., 2013). This produced an initial assembly consisting of 736.7 Mb sequence assembled and a contig N50 of 13.8 Mb. A combination of syntenic markers from BTx642 and primary annotated genes from BTx623 were used to identify misjoins in the assembly. Syntenic markers consisted of a total of 32,400 unique, non-repetitive, non-overlapping 1.0 Kb sequences generated using the version 1.0 S. bicolor BTx642 genome release. The gene set consisted of a subset of 12,641 BTx642 annotated primary genes from the V1 annotation. Both the syntenic markers and the primary gene set were aligned to the polished assembly and a total of 1 misjoins were identified in the assembly. Contigs were then ordered, oriented, and assembled into chromosomes using the combination of syntenic markers and genes. Contigs terminating in significant telomeric sequence were identified using the (TTTAGGG)n repeat, and care was taken to make sure that they were properly oriented in the release assembly. A total of 73 joins were applied to the assembly to form the final assembly consisting of 10 chromosomes. Each chromosome join is padded with 10,000 Ns. Chromosomes were numbered using the V1.0 S. bicolor BTx642 assembly available from Phytozome (Goodstein et al., 2012). The remaining scaffolds were screened against bacterial proteins, organelle sequences, GenBank nr and removed if found to be a contaminant. Chloroplast and mitochondrial genomes were generated using the OatK pipeline (Zhou et al., 2024). Homozygous SNPs and INDELs were corrected in the release using 49.7x of the Illumina fragment reads (2x150, 400bp insert) by aligning the reads using bwa mem (Li, 2013) and identifying homozygous SNPs and INDELs with the GATK's UnifiedGenotyper tool (McKenna et al., 2010). A total of 16 homozygous snps and 2,067 homozygous indels were corrected in the consensus sequence. The final release consisted of 707.2 Mb of sequence with a contig N50 of 25.4 Mb.

Supplementary Note 3 | Comparison of graph and single-reference genotyping

Here, we conducted a comparison of graph and single-reference genotyping approaches at a known presence-absence variation site (Supplementary Figure 2). To illustrate the issues, we chose a locus with copy number variation of phenylalanine ammonia lyase on chromosome 4 across the 33 pangenome members (Supplementary Figure 2A). The BTx623 reference and SC 971, which harbors the largest deletion of three are the top and bottom genomes and gene arrows indicate the presence and orientation of the bounding genes that all are single copy (lavender) or the CNV variable phenylalanine ammonia lyase gene members (dark purple). From left to right the six gene models in BTx623 are: a: Sobic.004G220300, b: Sobic.004G220400, c: Sobic.004G220500, d: Sobic.004G220600, e: Sobic.004G220650, and f: Sobic.004G220700. PI numbers (or genome IDs) from bottom to top: BTx623, PI656057, PI597980, PI656044, PI656027, PI565121, PI276837, PI513676, PI569459, RTx430, Wray, PI329301, PI276816, PI329501, PI656023, PI180348, PI534133, PI655981, PI660557, PI533766, PI660565, PI585966, PI570071, PI156178, PI660563, PI656031, PI656015, PI655988, PI576434, PI154844, BTx642, PI656050, PI656111. Positions are provided in the BTx623 coordinate system, as BTx623 is the primary reference in the pangenome graph. Positions of genes in other genomes are approximate. Gene PAV was inferred using odgi (Guarracino et al., 2022) untangle and a minimum Jaccard similarity of 0.8 with respect to BTx623 gene sequences.

Given the clear deletions in this region, it is likely that previously observed heterozygosity at SNPs could be due to methodological issues. To test this, we conducted a comparison of depth and genotype calls between linear and graph alignments of phenylalanine ammonia lyase array on sorghum chromosome 4 (Supplementary Figure 2B). To generate genotype calls, short reads were aligned to both the linear V5 BTx623 reference genome with bwa-mem (Li & Durbin, 2009) and the Minigraph-Cactus (Hickey et al., 2023) pangenome graph with 33 haplotypes and BTx623 as the primary reference (vg giraffe). Graph-aligned reads were 'surjected' to BTx623 coordinate space and an identical variant-calling workflow was applied to linear- and graph-based BAM files. Concordant with the PAV results, graph-based alignments produce missing calls for SC 971 at genes c-e, whereas linear-based alignments primarily produce homozygous reference genotype calls at these loci. Additionally, linear-based alignments produced heterozygous calls not observed in the graph-based calls across the region, likely due to the reference bias inherent in linear-based analyses.

Supplementary Note 4 | Choosing the optimal subpopulation *k* for ADMIXTURE

In most situations, the true number of genetic subpopulations (*k*) is not known a priori. Furthermore, the interaction of ongoing geneflow, vicariance, introgressions, and adaptation, may obscure the most biologically relevant *k*, even if the original number of subpopulations is known. In the case of sorghum, despite the hypothesis of between two and four independent domestications, and between three and seven botanical types, most previous papers suggest the optimal *k* is greater than eight. To determine the best *k* value, we first calculated the 'CV error' statistic for $k = 2$-13 (Supplementary Figure 3A). This statistic can sometimes be informative, where an optimal value shows a clear drop relative to the predicted value from a non-linear regression (NLS). In our case, the best values by this measure were $k = 4$, 5, 10, and 12. For main text figures and analyses, we opted to present $k = 10$ for two reasons (Figure 1). First, among the group of high *k* runs with low CV error values, $k = 10$ was the lowest *k* with the lowest respective CV error. That is, even though $k = 12$ and 13 had lower errors, these did not outweigh the increase in complexity that the higher k values would incur. Secondly, most recent analyses of sorghum diversity had $k \geq 8$. While $k = 4$ seems like a decent option, the CV error of that run (0.512) was 15% higher than the CV of $k = 10$ (0.446). When combined with the precedent of higher numbers of subpopulations in previous work, we opted to use $k = 10$ for all subsequent analyses. We have additionally generated barplots of the full range of $k = 2$ -13 (Supplementary Figure 3B). As in Fig. 1, * indicates whether a bar in the plot is from a polishing library for one of the reference genomes. No matter the *k*-value chosen, the reference genomes span all subpopulations. The bars (resequencing libraries) here are horizontally ordered by subpopulation ID, then by ancestry therein. Subpopulation IDs are not equivalent between *k*-values, so order from one *k* to the next is not consistent.

Supplementary Note 5 | Details regarding phenotype data collection and processing

Field sites and experimental design. Field experiments were conducted in 2023 and 2024 at two locations: 'MO', the Danforth Field Research Site (FRS) in St. Charles, Missouri (38.8472 N, −90.4493 W) and, 'SC' the Clemson University Pee Dee Research and Education Center in Florence, South Carolina (34.3072 N, −79.7447 W). A total of 255 sorghum accessions were grown each year in three replicate plots per accession at both sites. At the Missouri site, fields consist of a uniform DeSioux loam soil in a flood-free zone. The South Carolina site comprises Norfolk loamy sand (55%), Noboco loamy sand (28.9%), and Rains sandy loam (16.1%) soils (USDA-NRCS classifications).

Field measurements of flowering time and above-ground biomass. All field measurements were recorded in metric units, with designated check plots excluded from data collection. Flowering time (anthesis) was recorded in Missouri in 2023 and 2024. Beginning 60 days after planting, plots were surveyed twice weekly. For each plot, the date when ≥ 50 % of panicles had reached the flowering stage was recorded as the flowering date; plots below this threshold were left unscored until subsequent surveys. Each accession was represented by three independent field plots. Above-ground biomass was collected in both Missouri and South Carolina in 2023 and 2024 using the same experimental design (three replicate plots per accession per site). At maturity, two representative plants were harvested from each plot. Whole above-ground material was collected, dried to constant mass, and weighed to determine total above-ground biomass.

Sorghum Pangenome High-throughput Phenotyping Experiment (NS008). Three biological replicates of the sorghum pangenome reference lines were grown under two water regimes (totaling 210 plants). Seeds were planted in 8-inch tree pots containing 600 g of Turface® supplemented with Osmocote 14-14-14 fertilizer (1.5 lb per cubic yard). The potted seeds were held in a growth chamber at 32°C (14 h day) / 28°C (10 h night), 60% relative humidity, and 600 µmol $m^{-2}s^{-1}$ light intensity. After 6 days of germination, pots were transferred to the automated phenotyping platform at the Donald Danforth Plant Science Center's Bellwether Phenotyping Facility. Plants were maintained at 32°C (day) / 28°C (night) with a 14-hour light / 10-hour dark cycle, 60% relative humidity, and 600 µmol $m^{-2}s^{-1}$ light intensity. The system provided automated daily watering, weighing, and visible-light imaging. Plants were arranged in a randomized block design and rotated one-half lane daily to minimize edge effects. The soil was maintained at 100% field capacity (FC) by watering twice daily to a target mass of 1,267g, calculated as the sum of the carrier mass (344g), water at saturation (295g), and Turface-filled pot (628 g). Water added to reach the target mass was recorded for each pot at each weighing event. During the first four days, all plants were maintained at 100%FC; thereafter, the experiment was divided into two treatments: a drought-stress condition (40%FC) and a control (100%FC). Each plant was imaged daily using a visible light imaging chamber, capturing two side views and one top view per plant. After 28 days, plants were removed from the phenotyping system and shoot height and fresh weight were recorded immediately. Image analysis was performed using Plant Computer Vision (PlantCV), a Python-based toolkit, to extract plant area and height metrics from side-view images (Gehan et al., 2017). Water-use efficiency (WUE) was calculated as the ratio of plant projected area to cumulative water use: WUE (cm^2ml^{-1}) = Plant area / Cumulative water use. Plant area was derived from PlantCV segmentation data, and cumulative water use was obtained from daily irrigation weight records.

Physiological trait assessment (LI-600). Physiological traits were measured at the vegetative stage using a handheld LI-600 Porometer/Fluorometer (LI-COR Biosciences, Lincoln, NE, USA). Sampling was performed systematically across plots to capture genotypic variation. Measurements were taken on the youngest fully expanded leaf at a consistent canopy position to minimize within-plant variability. Recorded variables included ΦPSII (effective quantum yield of photosystem II; PhiPS2), leaf temperature (Tleaf), ambient light intensity (Qamb), leaf-level vapor pressure deficit (VPDleaf), and leaf angle. Data were collected between 09:00 and 13:00 local time to reduce heat-stress artifacts. The instrument was calibrated according to the manufacturer's guidelines prior to each session, and all readings were logged on the device and transferred to a central database for downstream analysis. This non-destructive protocol enabled rapid phenotyping of photosynthetic performance and water-use dynamics under field conditions.

UAV data collection and processing. UAV data collection and processing were conducted with the FieldDock platform, a fully autonomous, end-to-end system for high-throughput field phenotyping (fielddock.org). Before FieldDock's full deployment, the Taylor Geospatial Institute (TGI) supported data collection to maintain continuity: missions were manually planned and flown using flight software, with both RGB and multispectral sensors to capture high-resolution imagery. After FieldDock deployment in late 2024, operations were streamlined via a custom hexacopter equipped with a MicaSense multispectral camera. The UAV followed pre-programmed flight paths to ensure consistent acquisitions at the flowering–maturity stage in both years. Integrated onboard sensors and precise GPS calibration enabled accurate georeferencing for spatial analyses of crop traits. Post-flight imagery was processed within the FieldDock pipeline, including orthomosaic generation and vegetation-index computation, to extract agronomic features used in phenotypic analyses.

SUPPLEMENTARY REFERENCES

Cheng, H., Concepcion, G. T., Feng, X., Zhang, H., & Li, H. (2021). Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods*, *18*(2), 170–175. https://doi.org/10.1038/s41592-020-01056-5

Chin, C.-S., Alexander, D. H., Marks, P., Klammer, A. A., Drake, J., Heiner, C., Clum, A., Copeland, A., Huddleston, J., Eichler, E. E., Turner, S. W., & Korlach, J. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature Methods*, *10*(6), 563–569. https://doi.org/10.1038/nmeth.2474

Cole, B., Zhang, W., Shi, J., Wang, H., Baker, C., Varoquaux, N., Hollingsworth, J., Hutmacher, R., Dahlberg, J., Pierroz, G., Barry, K. W., Singan, V., Yoshinaga, Y., Daum, C., Zane, M., Blow, M., O'Malley, R., Shu, S., Jenkins, J. W., … Vogel, J. P. (2025). Multi-season analysis reveals hundreds of drought-responsive genes in sorghum. In *bioRxiv*. https://doi.org/10.1101/2025.06.27.662006

Cuevas, H. E., Zhou, C., Tang, H., Khadke, P. P., Das, S., Lin, Y.-R., Ge, Z., Clemente, T., Upadhyaya, H. D., Hash, C. T., & Paterson, A. H. (2016). The evolution of photoperiod-insensitive flowering in sorghum, A genomic model for panicoid grasses. *Molecular Biology and Evolution*, *33*(9), 2417–2428. https://doi.org/10.1093/molbev/msw120

Gehan, M. A., Fahlgren, N., Abbasi, A., Berry, J. C., Callen, S. T., Chavez, L., Doust, A. N., Feldman, M. J., Gilbert, K. B., Hodge, J. G., Hoyer, J. S., Lin, A., Liu, S., Lizárraga, C., Lorence, A., Miller, M., Platon, E., Tessman, M., & Sax, T. (2017). PlantCV v2: Image analysis software for high-throughput plant phenotyping. *PeerJ*, *5*, e4088. https://doi.org/10.7717/peerj.4088

Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N., & Rokhsar, D. S. (2012). Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Research*, *40*(Database issue), D1178–D1186. https://doi.org/10.1093/nar/gkr944

Guarracino, A., Heumos, S., Nahnsen, S., Prins, P., & Garrison, E. (2022). ODGI: understanding pangenome graphs. *Bioinformatics* , *38*(13), 3319–3326. https://doi.org/10.1093/bioinformatics/btac308

Hickey, G., Monlong, J., Ebler, J., Novak, A. M., Eizenga, J. M., Gao, Y., Human Pangenome Reference Consortium, Marschall, T., Li, H., & Paten, B. (2023). Pangenome graph construction from genome alignments with Minigraph-Cactus. *Nature Biotechnology*. https://doi.org/10.1038/s41587-023-01793-w

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., & Phillippy, A. M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research*, *27*(5), 722–736. https://doi.org/10.1101/gr.215087.116

Laursen, T., Borch, J., Knudsen, C., Bavishi, K., Torta, F., Martens, H. J., Silvestro, D., Hatzakis, N. S., Wenk, M. R., Dafforn, T. R., Olsen, C. E., Motawia, M. S., Hamberger, B., Møller, B. L., & Bassard, J.-E. (2016). Characterization of a dynamic metabolon producing the defense compound dhurrin in sorghum. *Science (New York, N.Y.)*, *354*(6314), 890–893. https://doi.org/10.1126/science.aag2347

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* , *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Lovell, J. T., Sreedasyam, A., Schranz, M. E., Wilson, M., Carlson, J. W., Harkess, A., Emms, D., Goodstein, D. M., & Schmutz, J. (2022). GENESPACE tracks regions of interest and gene copy number variation across multiple genomes. *eLife*, *11*. https://doi.org/10.7554/eLife.78526

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. https://doi.org/10.1101/gr.107524.110

Murphy, R. L., Klein, R. R., Morishige, D. T., Brady, J. A., Rooney, W. L., Miller, F. R., Dugas, D. V., Klein, P. E., & Mullet, J. E. (2011). Coincident light and clock regulation of pseudoresponse regulator protein 37 (PRR37) controls photoperiodic flowering in sorghum. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(39), 16469–16474. https://doi.org/10.1073/pnas.1106212108

Vaser, R., Sović, I., Nagarajan, N., & Šikić, M. (2017). Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Research*, *27*(5), 737–746. https://doi.org/10.1101/gr.214270.116

Xiao, C.-L., Chen, Y., Xie, S.-Q., Chen, K.-N., Wang, Y., Han, Y., Luo, F., & Xie, Z. (2017). MECAT: fast mapping, error correction, and de novo assembly for single-molecule sequencing reads. *Nature Methods*, *14*(11), 1072–1074. https://doi.org/10.1038/nmeth.4432

Zhou, C., Brown, M., Blaxter, M., The Darwin Tree of Life Project Consortium, McCarthy, S. A., & Durbin, R. (2024). Oatk: a de novo assembly tool for complex plant organelle genomes. In *bioRxiv*. https://doi.org/10.1101/2024.10.23.619857