scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Chromosome-level genome assemblies of sunflower oilseed and confectionery cultivars

Liuxi Yi[1,4], Haizhu Bao[1,4], Yang Wu[1,4], Yingnan Mu[2], Chao Du[3], Jingwen Peng[3], Xuechun Yan[3], Yongsheng Chen[1] ✉ & Haifeng Yu[2] ✉

The sunflower (*Helianthus annuus* L.), belonging to the Asteraceae family, is the world's fourth most important oil crop. Sunflower cultivars are categorized into oilseed and confectionery types. Here, we present chromosome-level genome assemblies of two Chinese sunflower cultivars-oilseed and confectionery-using PacBio HiFi and Hi-C sequencing. The oilseed cultivar, OXS, has a genome assembly spanning 3.03 Gb with 99.58% of sequences anchored to 17 chromosomes and a contig N50 length of 154.78 Mb. The first published confectionery cultivar genome, YDS, mirrors this closely with a 3.02 Gb assembly, contig N50 length of 153.87 Mb and 99.40% of sequences mapped similarly. Gene completeness reached 98.2% for OXS and 98.4% for YDS, with LTR Assembly Index scores of 24.73 and 25.85, respectively. Comparative genomics identified rapidly evolving gene families linked to synthesis, growth, and stress defense. Additionally, we found high collinearity between the YDS and OXS genomes, despite three significant inversions, and detected 15,056 large deletions and insertions. These findings lay a robust foundation for advanced genomic research and breeding innovations in sunflowers.

Background & Summary

The common sunflower (*Helianthus annuus L.*) is an important global oil crop belonging to the family Asteraceae. Originating from the Andes about 1600 years ago, sunflowers are now extensively cultivated worldwide due to their nutritional, medicinal, and environmental resilience, including tolerance to low temperatures, drought, and salt[1,2]. Sunflower has high nutritional, medicinal, and industrial value, with seed rich in unsaturated fats, proteins, vitamins, nutrients, phytosterols, tocopherols, and minerals. These seeds are utilized for oil extraction and in confectionery and bakery products[3–5]. Industrially, sunflower seeds and parts are used in pharmaceuticals[6] and as raw materials for organic dyes, cosmetics, margarine, plastics, perfumes, soaps, candles[7], and lubricants[8]. Sunflower is the world's fourth most important oil crop[9]. There are two main types of sunflower cultivars: oilseed and confectionery, with significant cultivation in Russia, Ukraine, the European Union, Argentina, and China. In the 2023/2024 season, global sunflower production reached 52.78 million metric tons recorded by USDA in their sunflowerseed production report (https://fas.usda.gov/data/production/commodity/2224000). Sunflower seeds contain about 40–45% oil, predominantly composed of linoleic and oleic acids, which make up around 90% of the oil's fatty acid content[10]. Sunflower oil has antioxidant properties that may reduce cardiovascular disease risk[11]. Sunflower kernels have a high protein content of 21% to 30%, although they are mainly used for the production of vegetable oil. Sunflower seeds processed as snacks are popular in countries like China, Eastern Europe, and the Middle East[12,13]. Compared to other nuts, confectionery sunflowers are affordable, nutritious, and convenient, with various flavors achieved through frying or boiling. Confectionery sunflower production accounts for 10% of total sunflower production. In China, 95% of the 0.60 million hectares of sunflower plantations are non-oilseed varieties used mainly for snacks, representing nearly half of the global consumption[14]. Confectionery sunflower is phenotypically distinct from oil sunflower in aspects such as plant

[1]Agricultural college, Inner Mongolia Agricultural University, Hohhot, Inner Mongolia, 010019, China. [2]Institute of Crop Science, Inner Mongolia Academy of Agricultural & Animal Husbandry Sciences, Hohhot, Inner Mongolia, 010031, China. [3]Bayan Nur Institute of Agriculture and Animal Husbandry Science, Bayan Nur, Inner Mongolia, 015000, China. [4]These authors contributed equally: Liuxi Yi, Haizhu Bao, Yang Wu. ✉e-mail: CYSimau@163.com; nkyyhf@163.com

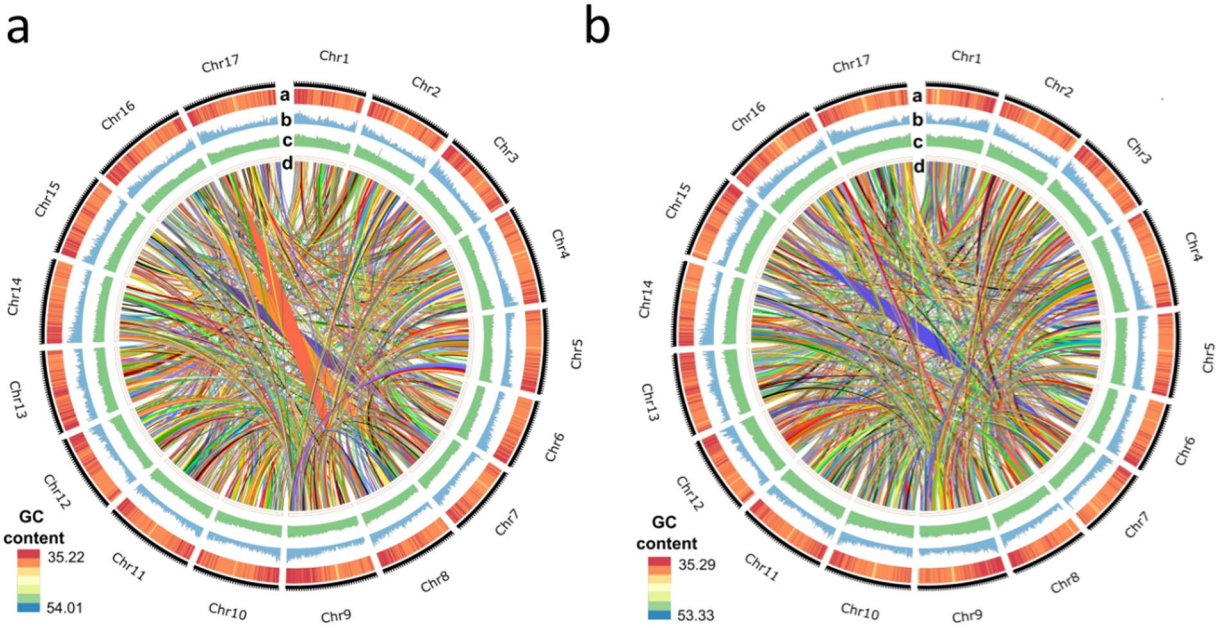

Fig. 1 Genomic characteristics of OXS and YDS. Genomic landscape of OXS (left, **a**) and YDS (right, **b**), including chromosome ideogram, GC content (**a**), gene density (**b**), repeat density (**c**), and intra-genome collinear blocks.

height, seed shape, and seed oil content. However, current research has mainly focused on oil seed lines, with few separate studies related to confectionery lines[15].

Sunflower (*Helianthus annuus*) karyotype analysis conducted by Otto Schrader *et al.* has confirmed that this species is diploid, possessing a chromosome number of 2n = 34[16]. Currently, nine sets of sunflower genomes have been made public, with genome sizes ranging from 3.00 G to 3.16 G and genome coverage between 10× and 172×. Contig N50 ranged from 635.3 Kb to 53.2 Mb[17]. These genomes were assembled to the chromosome level using the PacBio sequencing platform, with two using the PacBio RSII system and seven using the PacBio Sequel system. However, these genomes pertain mainly to oil-type domesticated and wild lines, with little research on non-oil types like confectionery sunflowers. There are no reports explaining the time of divergence and causes of phenotypic differences between oilseed and confectionery cultivars.

Here, we obtained high-quality genomes of two representative cultivars from oil and confectionery lineages in China at the chromosome level using the PacBio Revio system Circular Consensus Sequencing and high-throughput chromatin conformation capture (Hi-C) scaffolding sequencing technologies. OXS is an inbred oil-type sunflower line with high kernel rate (74.5% kernels), contains 42.9% oil and fat. It is highly susceptible to Verticillium wilt and moderately susceptible to several other diseases. YDS is an inbred non-oil line with high plant height (180–220 cm) and large plump seeds (19.23 grams per 100 seeds). The genome assembly of OXS, spans 3.03 Gb, with 99.58% of sequences anchored to 17 chromosomes and a contig N50 length of 154.78 Mb (Fig. 1). The Benchmarking Universal Single-Copy Orthologs (BUSCO) analysis yielded a gene completeness score of 98.2%. Repeat elements accounted for 80.77% of the genome, and 74,608 protein-coding genes were predicted. The complete circular mitochondrial (MT) and chloroplast (PT) genomes are 305,253 bp and 151,084 bp in length, respectively. Similarly, the assembly size of YDS, is 3.02 Gb, with 99.40% of sequences mapped to 17 chromosomes and a contig N50 length of 153.87 Mb (Fig. 1). The gene completeness of BUSCO reached 98.4%. Repeat elements accounted for 80.85% of the genome, and 73,244 protein-coding genes were predicted. The lengths of its MT and PT genomes are 305,259 bp and 151,098 bp, respectively. In this study, we found that the YDS genome assembly is highly collinear with the OXS assembly, except for three large inversions in Chr7 and Chr17. We detected 7,560 deletions, 7,496 insertions, and 2,034 breakends, which can provide a foundation for subsequent genetic improvement. The publication of these genomes significantly enhances genetic breeding efforts by integrating internal genetic and external environmental factors in *Helianthus annuus L.* crops.

Methods

Sample collection and sequencing. The *Helianthus annuus* L. samples used in this study was collected from the Tumochuan Plain in Inner Mongolia, China (111°40′91″ E, 40°50′90″ N). The OXS is an inbred oil-type sunflower line from Victory Inc, developed from the cross T-1063A × RT-039. This variety grows to a height of 178 cm and features purple stems. Its seeds weigh 5.7 grams per 100 seeds, and containing 74.5% kernels, 18% protein, and 42.9% oil and fat. The YDS is an inbred line of the JK601 hybrid, developed from the sterile line 30509 A and the restorer line 5 R by the Academy of Agricultural Sciences in Baicheng, China. This line grows to

a height of 180–220 cm, with a flower disk diameter of 19–22 cm. Its seeds containing 51.61% kernels, weighing 19.23 grams per 100 seeds, and containing 15.56% protein.

Fresh leaves were collected from each plant for genomic DNA (gDNA) extraction. Six different tissues from individual YDS plants and five from individual OXS plants (excluding leaves for OXS due to failed library construction) were collected for RNA extraction, including leaf, stem, root, seed, flower, and pollen. Library construction and sequencing were conducted at Novogene Co., Ltd (Tianjin, China). Genomic DNA was extracted and purified from leaves using Qiagen's MagAttract HMW DNA Kit (QIAGEN, Germantown, MD, USA), then sheared to a target size of 15–18 kb using the MegaRuptor 3 (Diagenode, Denville, NJ, USA), following the manufacturer's instructions. The HiFi sequencing library was prepared using the SMRTbell Express Template Prep Kit 2.0 (Pacific Biosciences, Menlo Park, CA, USA) and immediately treated with the Enzyme Clean Up Kit (Pacific Biosciences, Menlo Park, CA, USA). HiFi reads were generated using the CCS program v6.4.0 with settings: min-passes = 3, min-rq = 0.99 (https://ccs.how/). A Hi-C library was prepared by fixing DNA with formaldehyde, lysing the cells, digesting the DNA overnight with DpnII, biotinylating sticky ends, and ligating them to form chimeric junctions. These were enriched and sheared to 300–500 bp fragments, processed into paired-end sequencing libraries. Total RNA was extracted using Trizol reagent (Invitrogen, CA, USA) and purified using an NanoPhotometer® spectrophotometer (IMPLEN, CA, USA) following the manufacturer's protocol. RNA degradation and contamination were checked using 1% agarose gel electrophoresis and RNA integrity was assessed using an Agilent Bioanalyzer 2100 system (Agilent, CA, USA), with RIN values exceeding 8.5 for all samples. mRNA was enriched using Oligo (dT) beads, and sequencing libraries were prepared with an NEBNext® UltraTM RNA Library Prep Kit for Illumina® (NEB, USA), then sequenced on the Illumina Novoseq platform in 150 PE mode.

Genome assembly and assessment. The HiFi reads were assembled using Hifiasm 0.19.9-r616[18] and integrated with Hi-C reads to generate a primary and a pair of haplotype-resolved assemblies. Redundant haplotigs were removed using purge_dups v1.2.5[19] with specific parameters for OXS (-a 95, -T 5 32 32 33 33 144) and YDS (-T 5 31 31 32 32 144). Minimap2 v2.17-r941[20] was used to align the assembled contigs to the reference mitochondrial (NC_023337.1) and chloroplast (OR876284.1) sequences of *Helianthus annuus L*[21]. Contigs with 50% alignment were removed. Subsequently, the yahs 1.2a.1[22] pipeline (-e GATC) was used to anchor contigs onto chromosomes, followed by manual polishing using Juicebox Assembly Tools v2.20.00[23]. For the convenience of others in future comparisons and usage, the chromosomes were numbered and oriented according to the reference HanXRQr2.0-SUNRISE assembly. The Hi-C interaction heatmap was created utilizing HiCExplorer v3.6[24]. Using PMAT v1.5.3[25], the mitochondrial and chloroplast genomes were assembled from 10% of the HiFi reads. Bandage v0.9.0[26] was used to visualize non-nuclear genome graphs and to select an optimal path by connecting segments based on depth and strand alignment using Minimap2. We used QUAST v5.0.2[27] and BUSCO v5.7.1[28] for a quantitative assessment of genome assembly quality, employing the eudicots odb10 database in gene mode. In addition, the annotated and integrated proteins were evaluated using BUSCO with the eudicots odb10 database in protein mode. We used Minimap2 to align HiFi reads to the respective assemblies in order to assess the coverage of reads across the genome. The LTR Assembly Index (LAI) was used to evaluate the assembly continuity via the LTR_retriever pipeline.

Repeat element identification. We employed RepeatModeler v2.0.5[29] to built a custom repeat library, which identified TEs *de novo* using RECON v1.08[30] and RepeatScout v1.0.6[31]. Then, we used RepeatMasker v4.1.5[32] to identify repeat sequences in our genomes. High-quality long terminal repeat (LTR) families were discovered using LTRharvest 1.6.5[33] and LTR_retriever v2.9.5[34]. Finally, CD-HIT v4.8.1[35] was used to remove redundant TEs. The annotated and classified TE families were referenced against the Dfam v3.7[36] and Repbase version 23.08[37] databases.

Protein-coding genes prediction and function annotation. We employed Braker3 pipeline v3.0.8[38], which can use genome, protein data and RNA-seq data to automatically train and predict highly reliable genes with GeneMark-ETP and AUGUSTUS v3.5.0[39], to perform gene structure prediction. The process began with ProtHint[40] to align 3,510,742 green plant orthologous genes from OrthoDB v11.1[41] to the genome sequences, providing essential hints for BRAKER3. Concurrently, RNA-Seq data were aligned to the duplicate-masked genome using HISAT2 v 2.2.1[42]. Following this, GeneMark-ETP was trained using RNA-Seq alignments and homologous protein evidence. Subsequently, AUGUSTUS was trained and predicted gene structures using the same extrinsic information along with GeneMark-ETP results. The prediction of UTRs was based on GUSHR (https://github.com/Gaius-Augustus/GUSHR)[43], which utilized RNA-Seq coverage information. Finally, the results from AUGUSTUS and GeneMark-ETP were combined using TSEBRA v1.1.2.5[44], resulting in the final gene predictions. The protein hints were first filtered through src = E (evidence provided by transcriptome) or src = C (chained evidence, meaning all hints of a group can be incorporated into a single transcript). The predicted protein genes, derived either entirely from computational methods or partially supported by hints, were filtered using the Python script selectSupportedSubsets.py which implemented in Braker3. We used DIAMOND v2.0.11[45] with the parameters:–moresensitive -p 64 -e 1e-6 -max-hsps 1 -k 1 -f 6 to align the annotated genes with the NR database[46] (https://www.ncbi.nlm.nih.gov/refseq/about/nonredundantproteins/) and Swiss-Prot[47] databases. Additionally, we employed the eggnog-mapper online annotation pipeline[48] (http://eggnog-mapper.embl.de) to annotate the genes to the eggNOG 5.0 database[49]. The InterProScan v 5.55–88.0[50] procedure was used for PFAM database annotation[51].

Syntenic and structural variants analysis. We first utilized Minimap2 with the parameters: -x asm10 -a--cs -r2000 -t 10 -k 28 -f 20, to perform genome-genome alignments. The HanXRQr2.0-SUNRISE genome

Sequencing Clean Data(Gb)	Statistic	OXS		YDS
	PacBio HiFi reads	198.91(14.25 M reads)		181.71(12.73 M reads)
	Hi-C	195.02		193.39
	RNA-seq(stem, root, seed, flower, pollen and leaf)	9.35, 4.53, 6.74, 8.19, 6.43, NA		7.31, 7.12, 6.77, 5.36, 6.36, 6.93
Genome Assembly	Statistic	OXS	YDS	HanXRQr2.0-SUNRISE
	Genome size (Gb)	3.03	3.02	3.01
	Number of chromosomes	17	17	17
	Number of organelles	2	2	2
	Number of scaffolds	82	123	332
	Scaffold N50(Mb)	178. 40	177.86	176.49
	Scaffold L50	8	8	8
	Number of contigs	107	128	2,712
	Contig N50(Mb)	154.78	153. 87	2
	Contig L50	9	9	448
	GC percent(%)	38.70	38.71	38.59
	mtDNA size(Bp)	305,253	305,259	300,945
	ptDNA size(Bp)	151,084	151,098	151,104
LTR Assembly Index scores		24.73	25.85	12.65

Table 1. Assembly and assessment of OXS, YDS and HanXRQr2.0-SUNRISE genome assemblies.

BUSCO Assessment(gene mode)	Statistic	OXS	YDS	HanXRQr2.0-SUNRISE
	Complete BUSCOs (%)	2284(98.2%)	2287(98.4%)	2263(97.3%)
	Complete and single-copy BUSCOs (%)	1982(85.2%)	1983(85.3%)	1952(83.9%)
	Complete and duplicated BUSCOs (%)	302(13.0%)	304(13.1%)	311(13.4%)
	Fragmented BUSCOs (%)	15(0.6%)	13(0.6%)	16(0.7%)
	Missing BUSCOs (%)	27(1.2%)	26(1.0%)	47(2.0%)
	Total BUSCO groups searched	2326		
BUSCO Assessment(protein mode)	Complete BUSCOs (%)	2247(96.6%)	2245(96.5%)	2241(96.3%)
	Complete and single-copy BUSCOs (%)	1759(75.6%)	1720(73.9%)	1947(83.7%)
	Complete and duplicated BUSCOs (%)	488(21.0%)	525(22.6%)	294(12.6%)
	Fragmented BUSCOs (%)	4(0.2%)	11(0.5%)	25(1.1%)
	Missing BUSCOs (%)	75(3.2%)	70(3.0%)	60(2.6%)
	Total BUSCO groups searched	2326		

Table 2. BUSCO Assessment of OXS, YDS and HanXRQr2.0-SUNRISE genome assemblies.

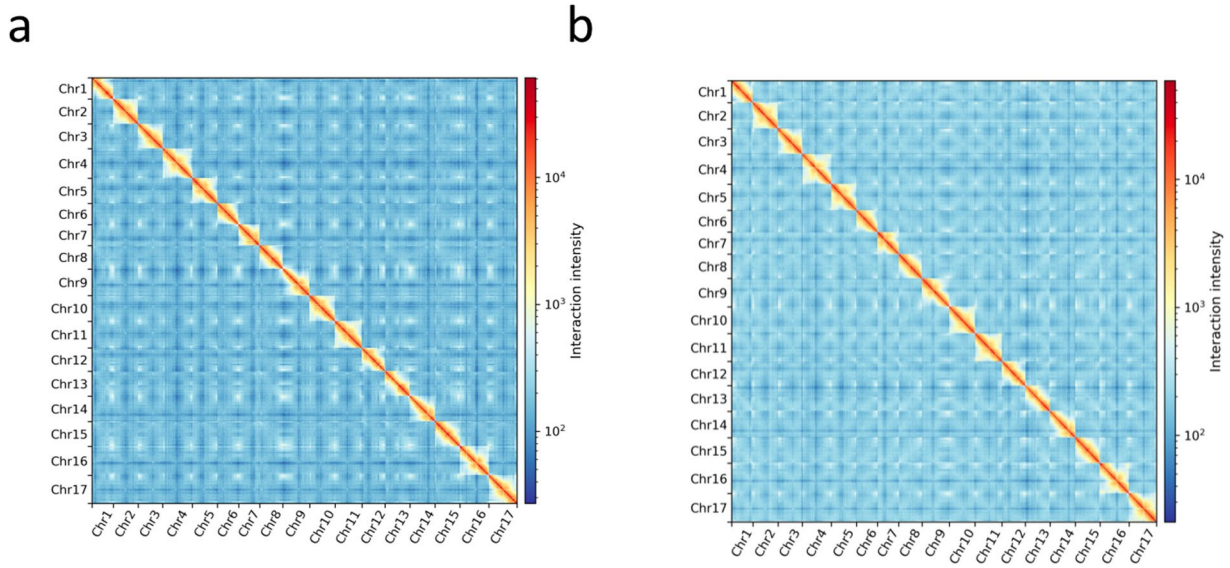

Fig. 2 Hi-C interaction heatmap for OXS (left, **a**) and YDS (right, **b**). The map shows scaffolded and independently assembled chromosomes at high resolution in 5 Mb windows.

Repeat elements	OXS			YDS		
	Number	Length (bp)	Proportion in genome (%)	Number	Length (bp)	Proportion in genome (%)
SINEs	873	81,858	0.00	2,204	205,383	0.01
LINEs	74,877	43,946,231	1.45	85,898	49,329,328	1.63
LTR elements	933,472	1,518,140,450	50.02	913,286	1,439,675,822	47.63
LTR-Ty1/Copia	229,776	277,293,055	9.14	232,332	314,713,904	10.41
LTR-Gypsy/DIRS1	653,149	1,196,934,249	39.44	651,712	1,107,371,339	36.63
DNA transposons	248,633	78,830,844	2.60	286,529	85,772,237	2.84
Unclassified	2,332,326	770,980,417	25.40	2,462,883	832,057,853	27.53
Total interspersed repeats(TEs)		2,411,979,800	79.47		2,407,040,623	79.63
Rolling-circles	15,453	8,828,544	0.29	1,7423	10,420,198	0.34
Small RNA	3,329	425,142	0.01	2,244	204,808	0.01
Satellites	18,390	5,457,066	0.18	6,509	1,702,665	0.06
Simple repeats	417,010	21,988,184	0.72	407,249	21,917,007	0.73
Low complexity	52,797	2,665,357	0.09	49,946	2,503,397	0.08
Total Tandem repeats	491,526	30,535,749	1.00	465,948	26,327,877	0.88
Total masked		2,451,303,894	80.77		2,443,767,654	80.85

Table 3. Repetitive elements and their proportions in OXS and YDS common sunflower.

Statistic	OXS	YDS	HanXRQr2.0-SUNRISE
Number of protein-coding genes	74,608	73,244	71,248
Number of protein-coding transcripts	79,249	78,211	71,257
Total length of protein-coding gene (bp)	171,368,941	170,977,831	213,129,599
Average length of protein-coding gene (bp)	2,296.92	2,334.36	2,991.38
Number of exons	295,879	295,400	290,645
Average Number of exons per gene	3.96	4.03	4.08
Total exons length (bp)	92,212,509	91,611,780	9,3845,249
Average length of exon (bp)	311.65	310.13	322.89

Table 4. Statistics of protein-coding genes in OXS, YDS and HanXRQr2.0-SUNRISE.

Database	OXS		YDS	
	Number	Percent(%)	Number	Percent(%)
eggNOG	44,278	59.35	43,901	59.94
KOG	19,656	26.35	19,596	26.75
GO	19,425	26.04	19,304	26.36
Swissprot	37,236	49.91	37,032	50.56
Interproscan	50,626	67.86	49,909	68.14
Pfam	46,215	61.94	45,648	62.32
NR	73,159	98.06	71,733	97.94
Total of aligned	73,182	98.09	71,754	97.97
Total of unaligned	1,426	1.91	1,490	2.03

Table 5. Alignment of the OXS and YDS genomes against functional and protein databases.

assembly was utilized as the reference sequence, while the other genome assembly served as queries for comparison against it. When comparing the OXS and YDS assemblies, the YDS assembly was used as the reference. Synteny and collinearity between the assemblies were visualized using dot plots, which were generated with the R script pafCoordsDotPlotly.R (https://github.com/tpoorten/dotPlotly). Next, we employed svim-asm 1.0.3[52] to identify structural variants in haploid mode. Structural variants are typically defined as genomic variants larger than 50 bp.

Data Records

The PacBio Revio system Circular Consensus Sequencing data, high-throughput chromatin conformation capture (Hi-C) scaffolding sequencing data and transcriptomic sequencing data were deposited in the NCBI database under SRA accession SRP523180[53]. The genome assemblies were deposited in GenBank under accession numbers JBGKEE000000000[54] and JBGKED000000000[55], and they are also available for download from the Zenodo database[56].

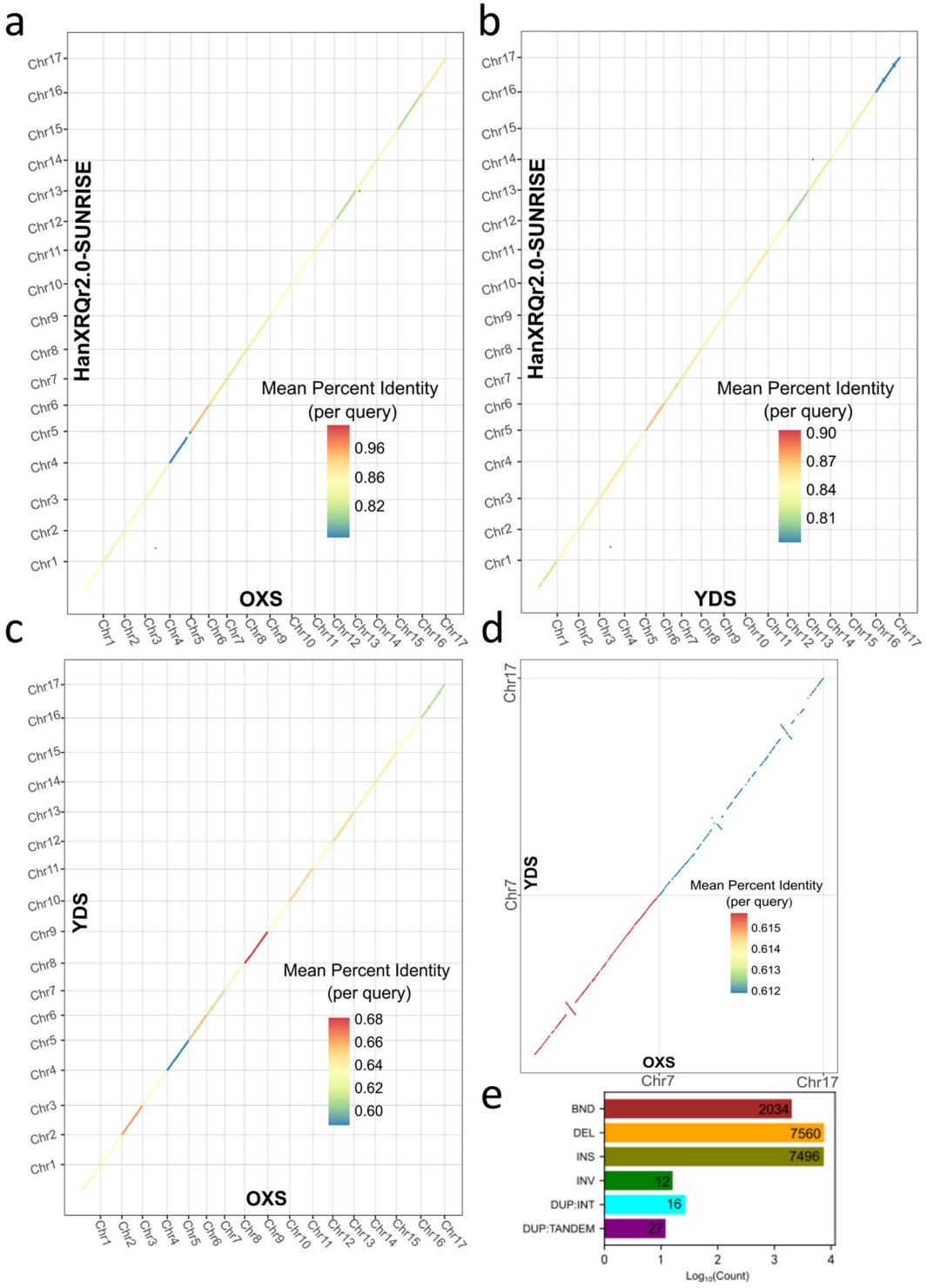

Fig. 3 Genome synteny map between the OXS, YDS and HanXRQr2.0-SUNRISE assemblies. (**a**) Genome synteny map between the OXS and HanXRQr2.0-SUNRISE. (**b**) Genome synteny map between the YDS and HanXRQr2.0-SUNRISE. (**c**) Genome synteny map between the YDS and OXS. (**d**) Three relatively large inversions of OXS and YDS genomes. (**e**) Statistics of various types of SVs between the OXS and YDS assemblies. BND: breakends, DEL: deletions, INS: insertions, INV: inversions, DUP:INT: interspersed duplications, DUP:TANDEM: tandem duplications. Annotation numbers are counts.

Technical Validation

Genome sequencing and assembly. A total of 198.9 Gb of HiFi data (OXS), including 14.25 M reads, with maximum lengths of 49,823 bp, average lengths of 13,955 bp and read length N50 of 14,083 bp and 181.7 Gb of HiFi data (12.74 M reads, YDS) were generated, with maximum lengths of 52,831 bp, average lengths of 14,267 bp,

and read length N50 of 14,029 bp (Table 1). For Hi-C sequencing, 195.02 Gb of high-quality data with a Q30 ratio of 95.63% for OXS and 193.40 Gb of Hi-C data with a Q30 ratio of 96.06% for YDS were obtained. For RNA-Seq, an average of 7.0 Gb per tissue of OXS with Q30 ranging from 95.56% to 96.60%, and an average of 6.6 Gb per tissue of YDS with Q30 ranging from 94.39% to 96.83%, were generated. The OXS primary assembly consisted of 1,548 contigs with a total length of 3.10 Gb and a contig N50 value of 146.41 Mb. The YDS primary assembly included 1,373 contigs, totaling 3.07 Gb with a contig N50 value of 153.87 Mb. Using the eudicots_odb10 database, we identified 98.2% and 98.3% of eudicot conserved single-copy homologous genes in the OXS and YDS primary assemblies, respectively (Table 2). The removal of redundant haplotigs resulted in 292 contigs with a total length of 3.04 Gb, a contig N50 value of 154.78 Mb, and 98.2% complete BUSCOs for OXS, and 187 contigs with a total length of 3.02 Gb, a contig N50 value of 153.87 Mb, and 98.4% complete BUSCOs for YDS. Using Hi-C data, scaffolds were successfully anchored to 17 pseudo-chromosomes, covering 99.58% of the total length for OXS and 99.40% for YDS. We ultimately achieved the final chromosome-scale genome assembly of OXS (3.03 Gb) with 82 scaffolds and a scaffold N50 value of 178.40 Mb, and YDS (3.02 Gb) with 123 scaffolds and a scaffold N50 value of 177.86 Mb (Table 1). The completeness of the BUSCO assessment remained high (98.2% for OXS, 98.4% for YDS) after Hi-C scaffolding, surpassing the 97.3% of the reference genome HanXRQr2.0-SUNRISE (Table 2). LTR Assembly Index scores reached 24.73 for OXS and 25.85 for YDS, respectively, whereas the HanXRQr2.0-SUNRISE achieved a score of 12.65 (Table 1). Furthermore, the complete BUSCO scores in protein mode for the two cultivars were 96.6% and 96.5%, respectively, indicating a high level of gene annotation quality (Table 2). These scores surpass the 96.3% achieved by the HanXRQr2.0-SUNRISE genome. The complete circular mitochondrial (MT) and chloroplast (PT) genomes of OXS were 305,253 bp and 151,084 bp, respectively. For YDS, the MT and PT genomes were 305,259 bp and 151,098 bp, respectively. The PT genome size was close to the published OR876284.1 genome (151,087 bp). By aligning HiFi reads to the respective assemblies, we achieved read mapping rates of 100% for both OXS and YDS. All genome assembly statistics and Hi-C interaction heatmap are detailed in Tables 1, 2 and Fig. 2.

Repeat element identification. The repeat element analysis revealed 2.45 Gb and 2.44 Gb of repetitive sequences, comprising 80.77% of the OXS genome and 80.85% of the YDS genome, respectively. In the OXS genome, we identified 2.41 Gb (79.47%) of TEs, with LTRs making up the majority of repeats (50.02%), predominantly Gypsy/DIRS1 elements (39.44%). Additionally, 30.53 Mb (1.0%) were tandem repeats, which included small RNA, satellite, simple repeats, and low complexity repeats. Similarly, in the YDS genome, we identified 2.40 Gb (79.63%) of TEs and 26.32 Mb (0.88%) of tandem repeats. LTRs also constituted the majority of repeats (47.63%), with Gypsy/DIRS1 elements (36.63%) being the most prevalent (Table 3).

Protein-coding genes prediction and function annotation. We predicted a total of 321,395 genes, containing 351,763 transcripts for OXS, and 304,410 protein-coding genes, containing 333,833 transcripts for YDS. After applying the filtering steps in Braker3, we obtained predictions of 74,608 protein genes and 79,249 corresponding transcripts for the OXS assembly, and 73,244 protein genes and 78,211 transcripts for the YDS assembly (Table 4). And the HanXRQr2.0-SUNRISE genome contains 71,248, 71,257 protein-coding genes and corresponding protein-coding transcripts. Then, we obtained comprehensive gene function annotations by utilizing several databases, including eggNOG, KEGG, GO, Swiss-Prot, and Pfam, as well as through homologous comparison with the NR database (Table 5). In the OXS genome, a total of 73,182 genes (98.09% of the total) successfully aligned with at least one database. Similarly, in the YDS genome, 71,754 genes (97.97% of the total) were successfully aligned.

Syntenic and structural variants analysis. Our analysis revealed that the confectionery cultivar genome (YDS) assembly is highly collinear with the oilseed cultivar assembly (OXS and HanXRQr2.0-SUNRISE) (Fig. 3a–c), except for three relatively large inversions located in Chr7 and Chr17 (Fig. 3d). Furthermore, we detected 7,560 deletions, 7,496 insertions, 2,034 breakends, 12 inversions, 16 interspersed duplications, and 27 tandem duplications when comparing the YDS and OXS genomes (Fig. 3e).

Code availability

In this study, all analyses were performed according to the manuals and tutorials provided for each software and pipeline. The software and code used are publicly accessible. If specific parameters were not mentioned, default parameters recommended by the developers were used.

Received: 31 July 2024; Accepted: 6 November 2024;
Published online: 07 January 2025

References

1. Connor, D. J. & Jones, T. R. Response of sunflower to strategies of irrigation II. Morphological and physiological responses to water stress. *Field Crops Res.* **12**, 91–103 (1985).
2. Forleo, M. B., Palmieri, N., Suardi, A., Coaloa, D. & Pari, L. The eco-efficiency of rapeseed and sunflower cultivation in Italy. Joining environmental and economic assessment. *J. Clean. Prod.* **172**, 3138–3153 (2018).
3. Lofgren, J. R. Sunflower for confectionery food, bird food, and pet food in *Schneiter Sunflower Technology and Production* Vol. 35 (ACSESS press, 1997).
4. Seiler, G. J., Qi, L. L. & Marek, L. F. Utilization of sunflower crop wild relatives for cultivated sunflower improvement. *Crop Sci.* **57**, 1083–1101 (2017).
5. Pal, D. Sunflower (*Helianthus annuus L.*) seeds in health and nutrition in *Nuts and Seeds in Health and Disease Prevention* 1097–1105 (Academic Press, 2011).

6. Nandha, R., Singh, H., Garg, K. & Rani, S. Therapeutic potential of sunflower seeds: an overview. *Int. J. Res. Dev. Pharm. Life Sci.* **3**, 967–972 (2014).
7. Nguyen, D. T. C. *et al*. The sunflower plant family for bioenergy, environmental remediation, nanotechnology, medicine, food and agriculture: a review. *Environ. Chem. Lett.* **19**, 3701–3726 (2021).
8. Ghosh, P. & Karmakar, G. Evaluation of sunflower oil as a multifunctional lubricating oil additive. *Int. J. Ind. Chem.* **5**, 7 (2014).
9. Grompone, M. A. Sunflower Oil. in *Vegetable Oils in Food Technology Composition, Properties and Uses* 137–167 (Blackwell Press, 2011).
10. Radanović, A., Miladinović, D., Cvejić, S., Jocković, M. & Jocić, S. Sunflower genetics from ancestors to modern hybrids-a review. *Genes* **9**, 528 (2018).
11. Premnath, A., Narayana, M., Ramakrishnan, C., Kuppusamy, S. & Chockalingam, V. Mapping quantitative trait loci controlling oil content, oleic acid and linoleic acid content in sunflower (*Helianthus annuus L.*). *Mol. Breed.* **36**, 106 (2016).
12. Yilmaz, M. I. *et al*. *Determining yield stability in confectionery sunflower*. (II International Conference On Agricultural, Biological And Life Scienceat: Edirne, Turkey 2020).
13. Hladni, N. *Present status and future prospects of global confectionery sunflower production*. (Proceedings, 19th International Sunflower Conference, 2016).
14. Feng, J., Jan, C.-C. & Seiler, G. Breeding, production, and supply chain of confection sunflower in China. *OCL* **29**, 11 (2022).
15. Kaya, Y. *confectionery sunflower production in turkey* https://www.researchgate.net/publication/267798816 (2004).
16. Schrader, O. & Ahne, R. Karyotype analysis of *Helianthus annuus* using Giemsa banding and fluorescence *in situ* hybridization. *Chromosome Res.* **5**, 451–456 (1997).
17. Huang, K. *et al*. The genomics of linkage drag in inbred lines of sunflower. *PNAS* **120**, e2205783119 (2022).
18. Chen, T. *et al*. The genome sequence archive family: toward explosive data growth and diverse data types. *Genomics Proteomics Bioinformatics* **19**, 578–583 (2021).
19. Guan, D. *et al*. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* **36**, 2896–2898 (2020).
20. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
21. Badouin, H. *et al*. The sunflower genome provides insights into oil metabolism, flowering and Asterid evolution. *Nature* **546**, 148–152 (2017).
22. Dudchenko, O. *et al*. *De novo* assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
23. Dudchenko, O. *et al*. The Juicebox Assembly Tools module facilitates *de novo* assembly of mammalian genomes with chromosome-length scaffolds for under $1000. Preprint at https://doi.org/10.1101/254797 (2018).
24. Wolff, J. *et al*. Galaxy HiCExplorer: a web server for reproducible Hi-C data analysis, quality control and visualization. *Nucleic Acids Res.* **46**, W11–W16 (2018).
25. Bi, C. *et al*. PMAT: an efficient plant mitogenome assembly toolkit using low-coverage HiFi sequencing data. *Hortic. Res.* **11**, uhae023 (2024).
26. Wick, R. R., Schultz, M. B., Zobel, J. & Holt, K. E. Bandage: interactive visualization of *de novo* genome assemblies. *Bioinformatics* **31**, 3350–3352 (2015).
27. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075 (2013).
28. Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A. & Zdobnov, E. M. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol. Biol. Evol.* **38**, 4647–4654 (2021).
29. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci.* **117**, 9451–9457 (2020).
30. Bao, Z. & Eddy, S. R. Automated *de novo* identification of repeat sequence families in sequenced genomes. *Genome Res.* **12**, 1269–1276 (2002).
31. Price, A. L., Jones, N. C. & Pevzner, P. A. *De novo* identification of repeat families in large genomes. *Bioinformatics.* **21**, i351–i358 (2005).
32. Tempel, S. *Using and Understanding RepeatMasker. in Mobile Genetic Elements* vol. 859 29–51 (Humana Press, 2012).
33. Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software for *de novo* detection of LTR retrotransposons. *BMC Bioinformatics* **9**, 18 (2008).
34. Ou, S. & Jiang, N. LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol.* **176**, 1410–1422 (2018).
35. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
36. Hubley, R. *et al*. The Dfam database of repetitive DNA families. *Nucleic Acids Res.* **44**, D81–D89 (2016).
37. Bao, W., Kojima, K. K. & Kohany, O. Repbase update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* **6**, 11 (2015).
38. Gabriel, L. *et al*. BRAKER3: fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA. *Genome Res.* **34**, 769–777 (2024).
39. Keller, O., Kollmar, M., Stanke, M. & Waack, S. A novel hybrid gene prediction method employing protein multiple sequence alignments. *Bioinformatics* **27**, 757–763 (2011).
40. Brůna, T., Lomsadze, A. & Borodovsky, M. GeneMark-EP+: eukaryotic gene prediction with self-training in the space of genes and proteins. *NAR Genomics Bioinforma.* **2**, lqaa026 (2020).
41. Zdobnov, E. M. *et al*. OrthoDB v9.1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. *Nucleic Acids Res.* **45**, D744–D749 (2017).
42. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
43. Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using native and syntenically mapped cDNA alignments to improve *de novo* gene finding. *Bioinformatics* **24**, 637–644 (2008).
44. Gabriel, L., Hoff, K. J., Brůna, T., Borodovsky, M. & Stanke, M. TSEBRA: transcript selector for BRAKER. *BMC Bioinformatics* **22**, 566 (2021).
45. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
46. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **35**, D61–D65 (2007).
47. Bairoch, A. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* **28**, 45–48 (2000).
48. Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Mol. Biol. Evol.* **38**, 5825–5829 (2021).
49. Huerta-Cepas, J. *et al*. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* **47**, D309–D314 (2019).
50. Jones, P. *et al*. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).

51. Mistry, J. *et al*. Pfam: The protein families database in 2021. *Nucleic Acids Res.* **49**, D412–D419 (2021).
52. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
53. *NCBI Sequence Read Archive*. https://identifiers.org/ncbi/insdc.sra:SRP523180 (2024).
54. Yi, L. *Helianthus annuus* cultivar OXS, whole genome shotgun sequencing project. *GenBank* https://identifiers.org/ncbi/insdc:JBGKEE000000000 (2024).
55. Yi, L. *Helianthus annuus* cultivar YDS, whole genome shotgun sequencing project. *Genbank* https://identifiers.org/ncbi/insdc:JBGKED000000000 (2024).
56. Yi, L. Chromosome-level genome assemblies of sunflower oilseed and confectionery cultivars. *Zenodo* https://doi.org/10.5281/zenodo.13123635 (2024).

Acknowledgements

This study was funded by the Inner Mongolia Natural Science Foundation (2024ZD19, 2023MS03008), operating expenses of basic scientific research projects of directly affiliated universities in Inner Mongolia (BR22-11-01, BR231513), and the National Modern Agricultural Industry Technology System funded by the Ministry of Finance and the Ministry of Agriculture and Rural Affairs (CARS-14-1-02).

Author contributions

Yongsheng Chen and Haifeng Yu conceptualized and initiated the project. Yingnan Mu, Chao Du, Jingwen Peng, and Xuechun Yan were responsible for sample collection. Liuxi Yi, Haizhu Bao, and Yang Wu performed the analysis, and all three contributed to writing and reviewing the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Y.C. or H.Y.