Improved assembly and annotation of the sesame genome

Mingcheng Wang1,2*, Jianwei Huang3, Song Liu3, Xiaofeng Liu4, Rui Li2,5, Junjia Luo4,5*, Zhixi Fu4,*

1Institute for Advanced Study, Chengdu University, No. 2025 Chengluo Road, Chengdu 610106, China
2Engineering Research Center of Sichuan-Tibet Traditional Medicinal Plant, Chengdu University, Chengdu 610106, China
3Berry Genomics Corporation, Beijing 100015, China
4College of Life Sciences, Sichuan Normal University, Chengdu 610101, China
5School of Food and Biological Engineering, Chengdu University, Chengdu 610106, China

*To whom correspondence should be addressed. Email: lu0573581@163.com (J.L.); fuzx2017@sicnu.edu.cn (Z.F.)

Abstract

Sesame (Sesamum indicum L.) is an important oilseed crop that produces abundant seed oil and has a pleasant flavor and high nutritional value. To date, several Illumina-based genome assemblies corresponding to different sesame genotypes have been published and widely used in genetic and genomic studies of sesame. However, these assemblies consistently showed low continuity with numerous gaps. Here, we reported a high-quality, reference-level sesame genome assembly by integrating PacBio high-fidelity sequencing and Hi-C technology. Our updated sesame assembly was 309.35 Mb in size with a high chromosome anchoring rate (97.54%) and contig N50 size (13.48 Mb), which were better than previously published genomes. We identified 162.38 Mb repetitive elements and 24,345 high-confidence protein-coding genes in the updated sesame assembly. Comparative genomic analysis showed that sesame shared an ancient whole-genome duplication event with two Lamiales species. A total of 2,782 genes were tandemly duplicated. We also identified several genes that were likely involved in fatty acid and triacylglycerol biosynthesis. Our improved sesame assembly and annotation will facilitate future genetic studies and genomics-assisted breeding of sesame.

Keywords: sesame; oilseed crop; genome assembly; PacBio high-fidelity sequencing; oil metabolism

1. Introduction

Sesame (Sesamum indicum L.) is an erect annual herb that belongs to the Pedaliaceae family and has been widely cultivated in tropical and subtropical areas of Asia, Africa, and South America. It is commonly called “the queen of oilseeds” due to its high seed oil content (~50% of its weight) and quality. Sesame seed oil is a popular edible oil due to its pleasant and mild taste, excellent antioxidant activity, and high nutritional value. Sesame seeds and oil have been widely used in medicinal, cosmetic, and industrial applications, and have been proposed as a valuable source of biodiesel fuel. Sesame is one of two oldest oilseed crops known to man (the other is the coconut). Due to long-term natural and human selection, sesame possesses a high degree of morphological and genetic diversity among cultivated varieties. Elucidating the genetic mechanisms underlying oil-related trait diversity among sesame varieties will benefit the genetic breeding and improvement of this economically important crop.

With the rapid development of sequencing technologies, the availability of high-quality reference genomes and genome-wide surveys have provided the opportunity to perform genomics-assisted crop breeding. Wang et al. first reported an Illumina-based assembly of the sesame genotype, Zhongzhi No. 13, which was 274 Mb in length and 85.3% was anchored to 16 pseudomolecules based on a genetic map. This reference genome was further improved by a high-density genetic map, where 13 pseudomolecules corresponded with the haploid chromosomal number of sesame (2n = 26). In addition to Zhongzhi No. 13, several Illumina-based genome assemblies corresponding to different sesame genotypes have also been published. These sesame genome sequences have been widely applied in the genetic investigation of various phenotypes by genome-wide association study (GWAS), QTL mapping, gene family analysis, and pan-genome analysis. However, the publicly available sesame genomes consistently show low continuity with numerous gaps, which may reduce the quality of annotations and lead to missing crucial genomic information.

In this study, we constructed an improved genome assembly of sesame using PacBio high-fidelity (HiFi) reads and Hi-C data. Based on this updated sesame assembly, we generated a high-confidence set of protein-coding genes. We also performed a comparative genomic analysis of sesame and several other sequenced species, as well as investigated the evolutionary history of the sesame genome. The high-quality genome assembly presented here will facilitate future genetic studies and the genomics-assisted breeding of sesame.

2. Materials and methods

2.1. Sample preparation and sequencing

Fresh leaves were harvested from adult S. indicum cv. Baizhima plants grown in Mianyang, Sichuan Province,
Southwest China. The variety Baizhima is quite different from other sequenced sesame varieties in several agronomic traits. It has a late flowering habit with increased branches and small white coat seeds. Total genomic DNA was extracted using the cetyl trimethylammonium bromide (CTAB) method for whole-genome DNA sequencing. First, Illumina ReSeq libraries were generated using previously described methods, including DNA extraction and digestion, DNA ligation, and purification. The Hi-C libraries were sequenced on an Illumina NovaSeq 6000 platform to generate Hi-C reads for the chromosome-level assembly.

In addition to DNA sequencing, total RNA was extracted from fresh leaf, stem, flower, seed, and husk tissues from the same individual sesame plant. RNA-seq libraries were prepared following the TruSeq Stranded mRNA-Seq protocol (Pacific Biosciences, Menlo Park, CA, USA). The resulting HiFi reads were used for the contig-level genome assembly. Finally, we collected ≥2 g young leaves and prepared Hi-C libraries following previously described methods, including chromatin extraction and digestion, DNA ligation, and purification. The Hi-C libraries were sequenced on an Illumina NovaSeq 6000 platform to generate Hi-C reads for the chromosome-level assembly.

Before assembly, the genome size of sesame was estimated based on a 19-mer frequency distribution of Illumina reads generated using Jellyfish v2.2.9. The homozygous peak depth of k-mer distribution was observed and the genome size was calculated using previously described methods with minor modifications. The HiFi reads were processed using CCS analysis workflow in SMRT Link v8.0 (Pacific Biosciences) with default parameters. To perform reassignment of allelic contigs caused by highly heterozygous regions, we identified and removed potential duplicate haplotypes from the contig-level assembly using Purge Haplotigs v1.1.2 with default parameters. The filtered HiFi reads were de novo assembled into contigs using hifiasm v0.14 with default parameters. To assign protein-coding regions of repetitive contigs caused by highly heterozygous regions, we used LTR-RTs.

2.4. Gene prediction
We predicted protein-coding genes within the repeat-masked v3 assembly using previously described methods. We aligned the protein sequences of Arabidopsis thaliana TAIR10, Solanum lycopersicum ITAG3.2, Mimulus guttatus v2.0, Utricularia gibba v1.0, Jacaranda mimosifolia v1.0, and Olea europaea v1.0 against our v3 assembly using BLAST. Predicted gene structures using GeneWise v2.4.19 were used to identify transcription factor (TF) genes.

In addition to protein-coding genes, we also annotated four types of non-coding RNAs (ncRNAs), including transfer RNA (tRNA), microRNA (miRNA), ribosomal RNA (rRNA), and small nuclear RNA (snRNA) based on structural features and homology searches. The miRNAs were predicted using miRNA-seq reads. The rRNAs were identified by aligning the publicly available rRNA sequences usingblastn.

2.5. Functional annotation of protein-coding genes
To assign gene functions, we first aligned the protein sequences against the Swiss-Prot and TrEMBL databases using DIAMOND v0.9.22. Conserved protein domains and motifs were identified using InterProScan v5.31. Gene ontology (GO) IDs were assigned to each gene using Blast2GO v2.5. The Kyoto Encyclopedia of Genes and Genomes (KEGG) Automatic Annotation Server was used to perform pathway mapping.
2.6. Gene family clustering and phylogenetic analysis
The protein sequences of sesame and five other sequenced plant species, J. mimosifolia, M. guttatus, O. europaea, S. lycopersicum, and A. thaliana, were used for the gene family analysis. These genes were clustered into families using OrthoFinder v2.3.11 with the DIMAOND aligner and Markov cluster (MCL) algorithm. Single-copy orthologs were extracted from the clustering results and multiple sequence alignment (MSA) was performed for each ortholog using MAFFT-LINSI v7.313. A maximum likelihood (ML) phylogenetic tree was constructed based on the concatenated alignments of all single-copy orthologs using RAxML v8.2.11 under the PROTGAMMAILGX model with 500 bootstrap replicates. Divergence times among the six plant species were estimated using MCMCTREE in PAML v4.9e with the following parameters: clock = 2, model = 3, and RootAge < 1.25. The divergence time between A. thaliana and S. lycopersicum (112–125 million years ago (Mya)) was obtained from the TimeTree database (http://www.timetree.org) and used as the fossil calibration point for the MCMCTREE analysis. Gene family expansions and contractions were determined for all internal nodes within the species tree using CAFE v3.1.

2.7. Whole-genome and tandem gene duplication analysis
To investigate the whole-genome duplication (WGD) history of sesame and related species, we identified syntenic blocks within and between genomes using MCScanX v1.1 based on the all-against-all pairwise comparisons of protein sequences of S. indicum, J. mimosifolia, M. guttatus, and O. europaea. For each ortholog and paralog gene pair, the nonsynonymous substitution rate (Ka) and synonymous substitution rate (Ks) were calculated using the “add_ka_and_ks_to_collinearity.pl” script in MCScanX. The duplication types were quantified using the “duplicate_gene_classifier” script.

We detected tandem gene duplications by searching homologous gene pairs that: (i) were located within 100 kb of each other; (ii) were located within 10 consecutive genes on a single scaffold; and (iii) had an identity > 50% and a coverage > 70% according to the BLASP results.

2.8. Analysis of genes involved in oil metabolism
The oil metabolism genes of sesame were identified by performing homology searches against Arabidopsis genes that are involved in fatty acid (FA) and triacylglycerol (TAG) biosynthesis. The expression levels of each gene in the seeds were evaluated using the Fragments per Kilobase per Million (FPKM) values calculated by Cufflinks v2.2.1. Genes with an FPKM value > 0.5 were considered to be expressed in the seeds.

3. Results and discussion
3.1. Improvement of the sesame genome assembly
A total of 24.33 Gb Illumina reads were generated for the genome survey. The genome size of sesame was estimated to be 317 Mb by the k-mer analysis (Supplementary Fig. S1), which was close to the genome size estimation (337 Mb) obtained by flow cytometry in a previous study. To construct a high-quality, reference-level sesame genome, we generated 34.33 Gb (108.29x genome coverage) HiFi reads with an average read length of 17.1 kb (Supplementary Table S1 and Fig. S2), which were de novo assembled into contigs. After discarding potential duplicate haplotypes, we retained 170 contigs with a total length of 309.34 Mb (Supplementary Table S2), covering 97.58% of the estimated genome size. These contigs were anchored to pseudochromosomes based on 32.34 Gb (102.00x genome coverage) Hi-C reads (Supplementary Table S1). In total, 301.73 Mb sequences (97.54% of the entire assembly) were assigned to 13 pseudochromosomes (Fig. 1A; Supplementary Table S3), all of which showed a well-organized diagonal pattern of intrachromosomal interactions (Supplementary Fig. S3). Compared to the v2 assembly, our v3 assembly showed a 13.42% increase in assembly size (309.35 Mb vs 272.73 Mb), a 288-fold reduction in the number of gaps (35 vs 10,076), and a 254-fold improvement in contiguity (contig N50: 13.48 Mb vs 53.07 kb) (Table 1 and Supplementary Table S2).

The high fidelity of our sesame v3 assembly was supported by the high mapping rate (99.57%) and 10-fold minimum genome coverage (99.13%) of the Illumina reads (Supplementary Table S4). The high completeness of our v3 assembly was also evidenced by the high mapping rate (94.56%) of the RNA-seq reads (Supplementary Table S5). Furthermore, we observed a BUSCO completeness score of 98.64% (Fig. 1B; Table 1) and an average LAI score of 19.22 for the entire assembly, which were better than the v2 assembly (BUSCO completeness score: 95.97%; LAI score: 13.92) (Fig. 1B; Supplementary Fig. S4). We found that 11 pseudochromosomes contained telomere sequences at both ends, and 2 pseudochromosomes contained telomere sequences at a single end, suggesting the good integrity of our v3 assembly (Supplementary Table S6). The MUMmer alignment analysis showed that the global collinearity was mostly conserved between the v2 and v3 assemblies (Supplementary Fig. S5) except for a large inversion corresponding to 8.48 Mb of genome sequence in “chr12” of the v3 assembly, indicating the high accuracy of the Hi-C scaffolding analysis. Several central regions of pseudochromosomes in the v3 assembly showed no similarity to any region of the v2 assembly and exhibited high levels of repetitive elements (Supplementary Table S7), indicating the advantage of PacBio sequencing in the assembly of highly repetitive regions. Collectively, these metrics indicated the high quality and completeness of our updated sesame assembly.

3.2. Updated annotation of the sesame genome
Using a combination of de novo and homology-based approaches, we identified 163.38 Mb repetitive elements representing 52.81% of the v3 assembly, which was much higher than the repeat content re-identified in the v2 assembly (109.80 Mb; 40.26% of the entire genome) (Table 1; Supplementary Table S8). Transposable elements (TEs) were the most abundant repeat class, spanning 84.84 Mb (27.42%) of the updated assembly (109.80 Mb) vs 53.07 kb) (Table 1; Supplementary Table S2). These contigs were anchored to pseudochromosomes based on 32.34 Gb (102.00x genome coverage) Hi-C reads (Supplementary Table S1). In total, 301.73 Mb sequences (97.54% of the entire assembly) were assigned to 13 pseudochromosomes (Fig. 1A; Supplementary Table S3), all of which showed a well-organized diagonal pattern of intrachromosomal interactions (Supplementary Fig. S3). Compared to the v2 assembly, our v3 assembly showed a 13.42% increase in assembly size (309.35 Mb vs 272.73 Mb), a 288-fold reduction in the number of gaps (35 vs 10,076), and a 254-fold improvement in contiguity (contig N50: 13.48 Mb vs 53.07 kb) (Table 1 and Supplementary Table S2).

The high fidelity of our sesame v3 assembly was supported by the high mapping rate (99.57%) and 10-fold minimum genome coverage (99.13%) of the Illumina reads (Supplementary Table S4). The high completeness of our v3 assembly was also evidenced by the high mapping rate (94.56%) of the RNA-seq reads (Supplementary Table S5). Furthermore, we observed a BUSCO completeness score of 98.64% (Fig. 1B; Table 1) and an average LAI score of 19.22 for the entire assembly, which were better than the v2 assembly (BUSCO completeness score: 95.97%; LAI score: 13.92) (Fig. 1B; Supplementary Fig. S4). We found that 11 pseudochromosomes contained telomere sequences at both ends, and 2 pseudochromosomes contained telomere sequences at a single end, suggesting the good integrity of our v3 assembly (Supplementary Table S6). The MUMmer alignment analysis showed that the global collinearity was mostly conserved between the v2 and v3 assemblies (Supplementary Fig. S5) except for a large inversion corresponding to 8.48 Mb of genome sequence in “chr12” of the v3 assembly, indicating the high accuracy of the Hi-C scaffolding analysis. Several central regions of pseudochromosomes in the v3 assembly showed no similarity to any region of the v2 assembly and exhibited high levels of repetitive elements (Supplementary Table S7), indicating the advantage of PacBio sequencing in the assembly of highly repetitive regions. Collectively, these metrics indicated the high quality and completeness of our updated sesame assembly.

3.2. Updated annotation of the sesame genome
Using a combination of de novo and homology-based approaches, we identified 163.38 Mb repetitive elements representing 52.81% of the v3 assembly, which was much higher than the repeat content re-identified in the v2 assembly (109.80 Mb; 40.26% of the entire genome) (Table 1; Supplementary Table S8). Transposable elements (TEs) were the most abundant repeat class, spanning 84.84 Mb (27.42%) of the updated assembly with LTR-RTs being the dominant type (44.36 Mb), followed by long interspersed nuclear elements (LINEs; 24.06 Mb), DNA transposons (16.28 Mb), and short interspersed nuclear elements (SINEs; 137.88 kb). Copia and Gypsy were the two most abundant LTR-RTs, representing 7.12% and 6.61% of the entire assembly, respectively. The divergence rate distribution analysis showed that most (70.83%) of the TEs had a divergence rate of < 20% (Supplementary Fig. S6). To gain further insight into the TE
amplification dynamics in sesame, we identified 778 intact LTR-RTs, including 417 Copia and 355 Gypsy elements. Most of the intact LTR-RTs (58.74%) were younger than 2 million years, with median insertion times of 1.03 and 1.52 Mya for the Copia and Gypsy elements, respectively (Supplementary Fig. S7). In addition to TEs, we identified 81.37 Mb (26.30% of the genome) unclassified repetitive sequences, 65.45% of which had a divergence rate of < 20% (Supplementary Fig. S8). Overall, the repetitive elements accumulated gradually with high recent activity within the sesame genome.

Combing de novo predictions, homology searches, and transcript evidence from the RNA-seq data, we identified 24,345 protein-coding genes, 23,947 (98.37%) of which were located on pseudochromosomes with an overall gene density of 79.4 gene per Mb (Supplementary Table S3). Overall, 21,916 (90.02%) protein-coding genes were functionally annotated by the Swiss-Prot and TrEMBL databases, 23,051 (94.68%) genes contained conserved domains recognized by the InterProScan analysis, 18,722 (76.90%) genes were assigned to GO terms, and 8,270 (33.97%) genes were mapped to KEGG pathways (Supplementary Table S9). Within the final gene set, we identified 1,223 TF genes belonging to 57 subfamilies (Supplementary Table S10). In addition to protein-coding genes, we identified 1,488 tRNAs, 129 miRNAs, 2,002 rRNAs and 350 snRNAs in the sesame genome (Supplementary Table S11).

Compared to the updated gene annotation set of the v1 assembly (hereafter referred to as v2 annotation), which was retrieved from the Ensembl Plants portal (http://www.plants.ensembl.org), our updated annotation set (hereafter referred to as v3 annotation) showed higher BUSCO completeness score (92.50% vs 81.78%) (Fig. 1B; Table 1), a longer average transcript length (3,422 bp vs 2,891 bp), similar average coding sequence (CDS) size (1,206 bp vs 1,207 bp), more exons per gene (5.98 vs 4.86), and a longer average intron length (445 bp vs 436 bp) (Table 1 and Fig. 1C). Among the 25,172 protein-coding genes in the v2 annotation set, 1,907 had no significant BLASTP hits to any annotated gene in our v3 annotation set, of which 1,534 (80.44%) did not have any functional annotation. These genes were significantly shorter than the other genes in the v2 annotation set (Wilcoxon rank sum test, p-value < 2.2e-16) (Supplementary Fig. S9). Therefore, we speculated that these genes may have been falsely annotated. Conversely, we identified 1,822 genes in our v3 annotation set that were missing from the v2 annotation set, of which 1,440 (79.03%) were supported by the functional annotation or gene expression analysis (Supplementary Fig. S10). Furthermore, we identified 13,973 1:1 orthologous gene pairs between the sesame v2 and v3 annotation sets by the MCScanX analysis, of which 7,138 were longer in our v3 annotation set than in the v2 annotation set, indicating that these genes may have been partially annotated in the previous assembly. Collectively, these findings indicated that our high-quality sesame assembly had more accurate structural annotations of the protein-coding genes.

3.3. Genome evolution of sesame
To investigate the evolutionary dynamics of the gene families in sesame, we clustered 22,044 (90.55%) sesame genes into 14,173 families by the OrthoFinder analysis (Fig. 2A). Among which, 9,798 (69.13%) families were shared by all six plant species and 334 families (containing 1,597 genes) were unique to sesame. The GO enrichment analysis of these sesame specific genes showed that they were highly enriched in “proton export across plasma membrane” (GO:0120029), “carbohydrate metabolic process” (GO:0005975), “regulation of...
### Table 1. Statistics of the assemblies and annotation sets of the sesame genome

| Assembly                        | v2                        | v3 (This study) |
|---------------------------------|---------------------------|-----------------|
| Sequencing platform             | Illumina Hiseq 2000       | PacBio Sequel II|
| Assembly size (bp)              | 272,734,981               | 309,354,609     |
| GC (%)                          | 34.98                     | 35.44           |
| Number of scaffolds             | 4,449                     | 135             |
| Scaffold N50 size (bp)          | 20,257,639                | 23,373,591      |
| Longest scaffold length (bp)    | 26,180,356                | 31,985,284      |
| Number of contigs               | 14,525                    | 170             |
| Contig N50 size (bp)            | 53,067                    | 13,482,096      |
| Longest contig length (bp)      | 471,223                   | 31,438,572      |
| Number of gaps                  | 10,076                    | 35              |
| BUSCO score (%)                 | 95.97                     | 98.64           |

| Annotation                      | v2                        | v3 (This study) |
|---------------------------------|---------------------------|-----------------|
| Repeat content (%)              | 40.26                     | 52.81           |
| Transposable elements (%)       | 21.59                     | 27.42           |
| Unclassified repeats (%)        | 18.66                     | 26.30           |
| Protein-coding genes            |                           |                 |
| Total number                    | 25,172                    | 24,345          |
| Average gene length (bp)        | 2,891                     | 3,422           |
| Average exon per gene           | 4.86                      | 5.98            |
| Average CDS length (bp)         | 1,207                     | 1,206           |
| Average intron length (bp)      | 436                       | 445             |
| BUSCO score (%)                 | 95.97                     | 98.64           |

### Conclusions

In this study, we reported an updated sesame assembly based on highly accurate, long-read HiFi sequencing and Hi-C sequencing data. Compared to the previously published Zhongzhi 13 assembly, our sesame v3 assembly had significantly higher contiguity and completeness. Based on this high-quality assembly, we predicted 24,345 high-confidence protein-coding genes with more accurate structural annotation. Evolutionary analysis showed that S. indicum and J. mimosifolia clustered together and these two species shared an ancient WGD event with M. guttatus. Additionally, we found that >10% of the protein-coding genes within the sesame genome were tandemly duplicated. We also identified several genes that were possibly involved in oil metabolism. Our high-quality, reference-level sesame assembly and improved genome annotations lay the foundation for elucidating the genetic basis of oil metabolism, abiotic stress resistance, and other key agronomic traits.

### Funding

This work was supported by National Natural Science Foundation of China (No. 32000158), Foundation of Sustainable Development Research Center of Resources and Environment of Western Sichuan, Sichuan Normal University (No. 2020CXY1HJZX03), the National Science & Technology Fundamental Resources Investigation Program of China (no. 2021XJJK0702), and the start-up funds provided by Chengdu University.

### Conflict of interest

The authors have no conflict of interest to declare.
Data Availability
The sesame v3 assembly and all sequence data are available in the NCBI database under the BioProject PRJNA875260. Genome assembly has been deposited at DDBJ/ENA/GenBank under the accession JAOXLQ00000000. RNA-seq data are available under the Sequence Read Archive (SRA) accession numbers SRX17378544-SRX17378548. Illumina short-read data, HiFi long-read data, and Hi-C data are available under the SRA accession numbers SRX17389771, SRX17602893, and SRX17602393. The updated genome assembly and annotations are also available on FigShare (https://doi.org/10.6084/m9.figshare.21151948).

Supplementary Data
Supplementary data are available at DNARES online.

References
1. Mujtaba, M.A., Cho, H.M., Masjuki, H.H., et al. 2020, Critical review on sesame seed oil and its methyl ester on cold flow and oxidation stability, Energy Rep., 6, 40–54.
2. Johnson, L.A., Suleiman, T.M., Lucas, E.W. 1979, Sesame protein: A review and prospectus, J. Am. Oil Chem. Soc., 56, 463–8.
3. Wan, Y., Li, H., Fu, G., et al. 2015, The relationship of antioxidant components and antioxidant activity of sesame seed oil, J. Sci. Food Agric., 95, 2571–8.
4. Namiki, M. 2007, Nutraceutical functions of sesame: a review, Crit. Rev. Food Sci. Nutr., 47, 651–73.
5. Ji, J., Liu, Y., Shi, L., Wang, N., Wang, X. 2019, Effect of roasting treatment on the chemical composition of sesame oil, LWT Food Sci. Technol., 101, 191–200.
6. Gharby, S., Harhar, H., Bouzoubaa, Z., et al. 2017, Chemical characterization and oxidative stability of seeds and oil of sesame grown in Morocco, J. Saudi Soc. Agric. Sci., 16, 105–11.
7. Saydut, A., Duz, M.Z., Kaya, C., Kafadar, A.B., Hamamci, C. 2008, Transesterified sesame (Sesamum indicum L.) seed oil as a biodiesel fuel, Bioresour. Technol., 99, 656–60.
8. Budowski, P., Markley, K.S. 1951, The chemical and physiological properties of sesame oil, Chem. Rev., 48, 125–51.
9. Pham, T.D., Thi Nguyen, T.D., Carlsson, A.S., Bui, T.M. 2010, Morphological evaluation of sesame (Sesamum indicum L.) varieties from different origins, Aust. J. Crop Sci., 4, 498–504.
10. Wei, X., Liu, K., Zhang, Y., et al. 2015, Genetic discovery for oil production and quality in sesame, Nat. Commun., 6, 1–10.
11. Yu, J., Golicz, A.A., Lu, K., et al. 2019, Insight into the evolution and functional characteristics of the pan-genome assembly from sesame landraces and modern cultivars, Plant Biotechnol. J., 17, 881–92.
12. Varshney, R.K., Bohra, A., Yu, J., et al. 2021, Designing future crops: genomics-assisted breeding comes of age, Trends Plant Sci., 26, 631–49.
Improved sesame assembly and annotation

13. Wang, L., Yu, S., Tong, C., et al. 2014, Genome sequencing of the high oil crop sesame provides insight into oil biosynthesis, Genome Biol., 15, R391–13.

14. Wang, L., Xia, Q., Zhang, Y., et al. 2016, Updated sesame genome assembly and fine mapping of plant height and seed coat color QTLs using a new high-density genetic map, BMC Genomics, 17, 1–13.

15. Wei, X., Zhu, X., Yu, J., et al. 2016, Identification of sesame genomic variations from genome comparison of landrace and variety, Front. Plant Sci., 7, 1169.

16. Li, D., Dossa, K., Zhang, Y., et al. 2018, GWAS uncovers differential genetic bases for drought and salt tolerances in sesame at the germination stage, Genes, 9, 87.

17. Asekova, S., Oh, E., Kulkarni, K.P., et al. 2021, An integrated approach of QTL mapping and genome-wide association analysis identifies candidate genes for phytophthora blight resistance in sesame (Sesamum indicum L.), Front. Plant Sci., 12, 604709.

18. Dossa, K., Diouf, D., Cissé, N. 2016, Genome-wide investigation of Hsf genes in sesame reveals their segmental duplication expansion and their active role in drought stress response, Front. Plant Sci., 7, 1522.

19. Mei, H., Liu, Y., Cui, C., et al. 2021, QTL mapping of yield-related traits in sesame, Mol. Breed., 41, 1–14.

20. Doyle, J.J., Doyle, J.L. 1987, A rapid DNA isolation procedure for small quantities of fresh leaf tissue, Phytoch. Bull., 19, 11–5.

21. Louwers, M., Splinter, E., Van Driel, R., De Laat, W., Stam, M. 2009, Studying physical chromatin interactions in plants using chromosome conformation capture (3C), Nat. Protoc., 4, 1216–29.

22. Bolger, A.M., Lohse, M., Usadel, B. 2014, Trimmomatic: a flexible trimmer for Illumina sequence data, Bioinformatics, 30, 2114–20.

23. Marçais, G., Kingsford, C. 2011, A fast, lock-free approach for efficient parallel counting of occurrences of k-mers, Bioinformatics, 27, 764–70.

24. Wang, M., Zhang, L., Tong, S., Jiang, D., Fu, Z. 2022, Chromosome-level genome assembly of a xerophytic plant, Haloxylon ammodendron, DNA Res., 29, dsac006.

25. Cheng, H., Concepcion, G.T., Feng, X., Zhang, H., Li, H. 2021, Haplotype-resolved de novo assembly using phased assembly graphs with hisfasm, Nat. Methods, 18, 170–5.

26. Roach, M.J., Schmidt, S.A., Borneman, A.R. 2018, Purge Haplottigs: allelic contig reassignment for third-gen diploid genome assemblies, BMC Bioinf., 19, 1–10.

27. Durand, N.C., Shamim, M.S., Machol, I., et al. 2016, Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments, Cell Syst., 3, 95–8.

28. Dudchenko, O., Batra, S.S., Omer, A.D., et al. 2017, De novo assembly of the AeDES aegypti genome using Hi-C yields chromosome-length scaffolds, Science, 356, 92–5.

29. Li, H., Durbin, R. 2009, Fast and accurate short read alignment with Burrows–Wheeler transform, Bioinformatics, 25, 1754–60.

30. Kim, D., Langmead, B., Salzberg, S.L. 2015, HISAT: a fast spliced aligner with low memory requirements, Nat. Methods, 12, 357–60.

31. Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., Zdobnov, E.M. 2015, BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs, Bioinformatics, 31, 3210–2.

32. Ou, S., Jiang, N. 2018, LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons, Plant Physiol., 176, 1410–22.

33. Riha, K., Shippen, D.E. 2003, Telomere structure, function and maintenance in Arabidopsis, Chromosoma, 11, 263–75.

34. Kurtz, S., Phillippy, A., Delcher, A.L., et al. 2004, Versatile and open software for comparing large genomes, Genome Biol., 5, 1–9.

35. Tarailo-Graovac, M., Chen, N. 2009, Using RepeatMasker to identify repetitive elements in genomic sequences, Curr. Protoc. Bioinformatics, 25, 4–10.

36. Price, A.L., Jones, N.C., Pevzner, P.A. 2005, De novo identification of repeat families in large genomes, Bioinformatics, 21, i351–8.

37. Jurka, J., Kapitonov, V.V., Pavlick, A., et al. 2005, Repbase update, a database of eukaryotic repetitive elements, Cytogenet Genome Res., 110, 462–7.

38. Xu, Z., Wang, H. 2007, LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons, Nucleic Acids Res., 35, W265–8.

39. Ellinghaus, D., Kurtz, S., Willhoft, U. 2008, LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons, BMC Bioinf., 9, 18.

40. Wang, M., Gu, Z., Fu, Z., Jiang, D. 2021, High-quality genome assembly of an important biodiesel plant, Erucaria lyrata L, DNA Res., 28, dsab022.

41. Wang, M., Tong, S., Ma, T., Xi, Z., Liu, J. 2021, The chromosome-level genome assembly of Sichuan pepper provides insights into apomixis, drought tolerance, and alkaloid biosynthesis, Mol. Ecol. Resour., 21, 2533–45.

42. Arabidopsis Genome Initiative. 2000, Analysis of the genome sequence of the flowering plant Arabidopsis thaliana, Nature, 408, 796–815.

43. Tomato Genome Consortium. 2012, The tomato genome sequence provides insights into frizzy fruit evolution, Nature, 485, 635–41.

44. Hellsten, U., Wright, K.M., Jenkins, J., et al. 2013, Fine-scale variation in meiotic recombination in Mimulus inferred from population shotgun sequencing, Proc. Natl. Acad. Sci. USA, 110, 19478–82.

45. Lan, T., Renner, T., Ibarra-Laclette, E., et al. 2017, Long-read sequencing uncovers the adaptive topography of a carnivorous plant genome, Proc. Natl. Acad. Sci. USA, 114, E4345–41.

46. Wang, M., Zhang, L., Wang, Z. 2021, Chromosomal-level reference genome of the neotropical tree Jacaranda mimosaefolia D. Don., Genome Biol. Evol., 13, evab094.

47. Unver, T., Wu, Z., Sterck, L., et al. 2017, Genome of wild olive and the evolution of oil biosynthesis, Proc. Natl. Acad. Sci. USA, 114, E9413–22.

48. Camacho, C., Coulouris, G., Avagyan, V., et al. 2009, BLAST+: architecture and applications, BMC Bioinf., 10, 421.

49. Birney, E., Clamp, M., Durbin, R. 2004, GeneWise and genomewise, Genome Res., 14, 988–95.

50. Haas, B.J., Delcher, A.L., Mount, S.M., et al. 2003, Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies, Nucleic Acids Res., 31, 5654–66.

51. Stanke, M., Keller, O., Gunduz, I., et al. 2006, AUGUSTUS: ab initio prediction of alternative transcripts, Nucleic Acids Res., 34, W435–9.

52. Haas, B.J., Salzberg, S.L., Zhu, W., et al. 2008, Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments, Genome Biol., 9, R7.

53. Jin, J., Tian, F., Yang, D.C., et al. 2016, PlantTDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants, Nucleic Acids Res., 45, D1040–5.

54. Lowe, T.M., Eddy, S.R. 1997, tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence, Nucleic Acids Res., 25, 955–64.

55. Nawrocki, E.P., Eddy, S.R. 2013, Infernal 1.1: 100-fold faster RNA homology searches, Bioinformatics, 29, 2933–5.

56. Bairoch, A., Apweiler, R. 2000, The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000, Nucleic Acids Res., 28, 45–8.
64. De Bie, T., Cristianini, N., Demuth, J.P., Hahn, M.W. 2006, CAFE: a computational tool for the study of gene family evolution, *Bioinformatics*, 22, 1269–71.

65. Wang, Y., Tang, H., DeBarry, J.D., et al. 2012, MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity, *Nucleic Acids Res.*, 40, e49.

66. Beisson, F., Koo, A.J., Ruuska, S., et al. 2003, Arabidopsis genes involved in acyl lipid metabolism. A 2003 census of the candidates, a study of the distribution of expressed sequence tags in organs, and a web-based database, *Plant Physiol.*, 132, 681–97.

67. Li-Beisson, Y., Shorrosh, B., Beisson, F., et al. 2013, Acyl-lipid metabolism, *Arabidopsis Book*, 11, e0161.

68. Trapnell, C., Roberts, A., Goff, L., et al. 2012, Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks, *Nat. Protoc.*, 7, 562–78.