# Improved chromosome-level genome assembly for marigold (*Tagetes erecta*)

Fan Jiang[1]*, Lihua Yuan[1,2,3]*, Sen Wang[1], Hengchao Wang[1], Dong Xu[1], Anqi Wang[1], and Wei Fan[1]†

**Affiliations:**

[1]Shenzhen Branch, Guangdong Laboratory for Lingnan Modern Agriculture; Genome Analysis Laboratory of the Ministry of Agriculture and Rural Affairs; Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, Guangdong, 518120, China.

[2]State Key Laboratory of Crop Stress Adaptation and Improvement, School of Life Sciences, Henan University, Kaifeng 475004, China.

[3]Shenzhen Research Institute of Henan university, Shenzhen 518000, China.

*These authors contributed equally to this work.

†Corresponding authors: Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen 518120, China. Email: fanwei@caas.cn

## Abstract

Marigold (*Tagetes erecta* L.) is a popular ornamental plant of the Asteraceae family, and its petals are considered the most abundant source of lutein. A low-continuity chromosome-level genome sequence of marigold was published recently, with poor annotation of the protein-coding genes, which hinders the studies of lutein biosynthesis. Here, we generated a near telomere-to-telomere level genome assembly of marigold based on highly accurate high-fidelity (HiFi) long reads and Hi-C sequencing data. Compared to the previously reported marigold genome, the current assembly had obviously higher contiguity and higher completeness of gene set. The current genome assembly has a 27-fold increase in contig N50 size, a 12.1% increase in chromosome anchoring rate, and a 9.0% increase in BUSCO complete rate for the gene set. Besides, the current assembly has much fewer assembly errors. Based on this high-quality genome assembly, we found that the 170-bp repeats are the most abundant centromeric unit and all centromeric regions are distributed along the whole chromosomes for all 12 centromeres, indicating the existence of the holocentromeres in marigold. In addition, we analyzed the structure and phylogenetic relationship of the four *PSY* genes, and revealed that these genes have diversified and possibly executed different functions in various tissues. Our near telomere-to-telomere level genome assembly and comprehensive gene annotation will greatly facilitate the breeding of marigold and researches aimed at improving lutein production.



**Keywords:** *Tagetes erecta*; marigold; Asteraceae family; near telomere-to-telomere assembly; PacBio HiFi sequencing


## 1. Introduction

44 Marigold (*Tagetes erecta* L.) is one of the most popular ornamental plants of the Asteraceae family,
45 and its flower petals has become the most important raw material for lutein production (Lin et al.,
46 2015; Zhang et al., 2020b; Zhang et al., 2020a). Marigold is an annual herbaceous plant with yellow
47 or orange flowers (Burlec et al., 2021; Moliner et al., 2018), and its flower petals are considered the
48 most abundant source of lutein (Nwachukwu et al., 2016; Park et al., 2017; Tanaka et al., 2008),
49 which is effective in the prevention and treatment of cancer, coronary artery disease, and many eye
50 diseases (Buscemi et al., 2018; Chung et al., 2017; Gansukh et al., 2019; Li et al., 2020).
51
52 The lutein biosynthetic pathway has been well studied in many high plants, and it starts with the
53 condensation of two geranylgeranyl diphosphate (GGPP) molecules to produce phytoene, which is
54 catalyzed by phytoene synthase (PSY) (Nisar et al., 2015; Sun et al., 2018). Subsequently, phytoene
55 is transformed to lycopene by a series of desaturation and isomerization reactions (Giuliano, 2014).
56 Then, lycopene is cyclized twice by lycopene $\varepsilon$-cyclase and lycopene $\beta$-cyclase to generate $\alpha$-
57 carotene, which is hydroxylated to lutein by hydroxylase (Kim et al., 2009). In this pathway, PSY
58 catalyzes the first and rate-limiting step in lutein biosynthesis (Maass et al., 2009; Nisar et al., 2015;
59 Rodriguez-Villalon et al., 2009), and four *PSY* genes in marigold were identified on the basis of a
60 whole genome sequence (Xin et al., 2023). However, the functional differences of these genes
61 among roots, stems, leaves, and flowers in marigold was less understood (Feng et al., 2018).
62
63 A recently study by Xin et al has reported a low-continuity marigold genome assembly, which
64 contains 3,451 contigs with a contig N50 size of 1.4 Mb, and only 87.6% of contig sequences were
65 anchored to 12 pseudo-chromosomes (Xin et al., 2023). In addition, the ratio of transposable

66 elements (57.3%) and BUSCO complete rate (89.3%) for the genes were obviously lower than those
67 of the closely related species (Badouin et al., 2017; Fan et al., 2022; Liu et al., 2020; Xu et al., 2021).
68 In this study, we generated a near telomere-to-telomere level reference genome assembly for
69 marigold, explored the structure of centromeres, and investigated the expression patterns of the *PSY*
70 genes in different tissues.
71

72 # 2. Materials and Methods

73 ## 2.1. Plant materials and genome sequencing

74 The seeds of marigold were provided by Suqian Sunrise Seed Industry Co., Ltd, in March 2022, and
75 then were grown in a greenhouse of Agricultural Genomics Institute at Shenzhen, Chinese Academy
76 of Agricultural Sciences, Guangdong province, China. Total genomic DNA was extracted from
77 fresh leaves using the Hi-DNAsecure Plant Kit (TIANGEN, China). Then, sequencing libraries
78 were constructed by SMRTbell Express Template Prep Kit 2.0 (PacBio, USA), and sequenced on a
79 PacBio Sequel II platform with Circular consensus sequence (CCS) mode (PacBio, USA). For Hi-
80 C sequencing, the cross-linked DNA sample was digested with the MboI at motifs GATC, and
81 paired-end sequenced (2 x 150 bp) on an Illumina NovaSeq 6000 platform.
82

83 ## 2.2. Transcriptome sequencing

84 To assist the prediction of protein-coding genes, total RNA samples were extracted by RNeasy Plant
85 Mini Kit (QIAGEN, Germany) from the different tissues, and then were reverse-transcribed into
86 cDNA using NEBNext Single Cell/Low Input cDNA Synthesis & Amplification Module (NEB,

87 UK). Sequencing libraries were constructed by SMRTbell Express Template Prep Kit 2.0 (PacBio,
88 USA) and sequenced on PacBio Sequel II system (PacBio, USA).
89
90 To investigate the expression patterns of candidate genes involved in lutein biosynthesis, the
91 samples collected from the roots, stems, leaves and flowers of 80-day-old plants were used for tissue
92 specificity analysis. Each plant represents a biological repeat, and five biological replicates were
93 used in this study. All samples were immediately frozen in liquid nitrogen and stored at -80°C until
94 use. Total RNA for each sample was extracted using RNeasy Plant Mini Kit (QIAGEN, Germany)
95 according to the manufacturer's instructions, and the cDNA libraries were paired-end (2 x 150 bp)
96 sequenced on an Illumina NovaSeq 6000 system.
97
98 **2.3 Genome assembly**
99 Before genome assembly, the PacBio HiFi reads with read length < 2,000 bp or average quality <
100 99% were filtered out. Then, the genome size was estimated by GCE v1.02 (Liu et al., 2013) based
101 on k-mer frequencies (K = 17). The filtered PacBio HiFi reads were assembled using Hifiasm
102 v0.16.1 (Cheng et al., 2021) with parameter "-l 3". Purge_dups v1.2.5 (Guan et al., 2020) with
103 parameters "-a 83 -f 0.8 -e" was used to filter the duplicate contigs in the primary assembly. After
104 removing contaminant contigs, we obtained 47 contigs with a total length of 778.9 Mb and a contig
105 N50 size of 38.0 Mb (**Table 1**). The completeness of the assembled contigs were evaluated using
106 the BUSCO v5.2.2 (Simao et al., 2015) with embryophyta_odb10 database, and the contig
107 correctness was estimated based on yak QV (https://github.com/lh3/yak).
108

The pseudo-chromosomes were constructed using Hi-C reads. We first mapped the Hi-C reads onto the assembled contigs by Bowtie2 v2.3.4.3 (Langmead and Salzberg, 2012), and then used HiC-Pro v2.11.4 (Servant et al., 2015) pipeline to detect valid ligation pairs and generate the Hi-C link matrixes among different contigs. At last, the contigs were clustered, ordered, and oriented into pseudo-chromosomes using EndHiC v1.0 (Wang et al., 2022) with parameters "--rounds 3 --binnumstep 6".

## 2.4. Transposable elements identification

Transposable elements (TEs) were identified through three steps: (1) we used EDTA v1.9.9 (Ou et al., 2019) to produce a filtered TE library; (2), we performed homology-searching against the above structural TE library, Repbase v26.05 (Bao et al., 2015), and protein-coding TE database using RepeatMasker v4.1.2 (Smit et al.); (3) we masked the above identified TEs with Ns in the genome, and used RepeatModeler v2.0.2 (Flynn et al., 2020) to construct an extra *de novo* TE library, which was further classified by TERL (da Cruz et al., 2021). All classified TE sequences in the *de novo* TE library were used by RepeatMasker to identify the remaining TEs in the genome.

## 2.5. Gene prediction and functional annotation

We used Augustus v3.4.0 (Stanke et al., 2006) to predict the protein-coding gene models. The training parameters for Augustus were generated during the BUSCO (Simao et al., 2015) completeness assessment of the genome assembly. The homology hints were generated by aligning the protein-coding sequences from the published genomes of *Smallanthus sonchifolius* (Fan et al., 2022), *Mikania micrantha* (Liu et al., 2020), *Helianthus annuus* (Badouin et al., 2017), and *Stevia*

131 *rebaudiana* (Xu et al., 2021) to marigold genome, using Exonerate v2.2.0 (Slater and Birney, 2005).
132 To produce transcription hints, the full-length transcripts and Illumina RNA-Seq reads were aligned
133 to the genome, and then were transformed to hints by blat2hints and bam2hints in Augustus,
134 respectively. The tRNA genes were predicted by tRNAscan-SE v2.0 (Chan et al., 2021), and the
135 rRNA genes were predicted by cmscan from infernal v1.1.4 (Nawrocki and Eddy, 2013). Functional
136 annotation was performed by searching the protein sequences against the NCBI-NR and KEGG
137 databases using DIAMOND v2 (Buchfink et al., 2021). The protein domain annotation was
138 performed using InterProScan v5.52-86 (Blum et al., 2021) against InterPro database.
139

140 **2.6. Identification of telomeres and centromeres**

141 To investigate the telomeres and centromeres, tandem repeats were explored using Tandem Repeats
142 Finder (TRF) v4.07 (Benson, 1999) with parameter "2 5 7 80 10 50 2000 -h -d". We identified
143 telomeres on the chromosome scaffolds based on TRF results and telomeric units of "TTTAGGG"
144 and "CCCTAAA". For centromere annotation, we investigated the tandem repeat units with length
145 of 100 - 200 bp, and counted the total length of each repeat units, revealing the 170-bp repeats were
146 the most abundant unit in the genome, followed by 167-bp, 178-bp and 169-bp repeats. The
147 distribution of the four repeat units in different pseudo-chromosomes were visualized by IGV
148 v2.11.4 (Thorvaldsdottir et al., 2013).
149

150 **2.7. Analysis of genes involved in lutein biosynthesis**

151 The lutein biosynthesis genes in marigold genome were identified by homology alignment. Firstly,
152 we downloaded the known genes involved in the lutein biosynthesis pathway (map00906) from

153 KEGG database for the closely related species in Asterales order, including *Helianthus annuus*,
154 *Erigeron canadensis*, *Lactuca sativa*, and *Cynara cardunculus* var. *scolymus*. Then, we aligned the
155 protein-coding genes of marigold to the downloaded lutein biosynthesis genes using DIAMOND
156 v2 (Buchfink et al., 2021) with parameters "blastp --more-sensitive --evalue 0.00001". The genes
157 in marigold with best alignment identity > 80% and coverage > 80% were taken as potential lutein
158 synthesis genes. Furthermore, the orthogroups constructed by OrthoFinder v2.5.2 (Emms and Kelly,
159 2019) that contained these genes were obtained and checked.
160
161 To construct the tree for *PSY* genes, we downloaded the protein sequences of *PSY* genes from the
162 KEGG database for the 4 Asterales species (*Helianthus annuus*, *Erigeron canadensis*, *Lactuca*
163 *sativa*, *Cynara cardunculus* var. *scolymus*, and *Solanum lycopersicum*) and 4 Solanales species
164 (*Solanum lycopersicum*, *Solanum pennellii*, *Solanum stenotomum*, and *Ipomoea triloba*). The tree
165 was constructed by fasttree v2.1.11 (Price et al., 2010) with parameters "-lg -gamma" based on
166 multiple sequences alignments from protein sequences. The expression level of these candidate
167 genes was evaluated by transcripts per million (TPM) calculated from the HISAT v2.2.1 (Kim et
168 al., 2019) mapping results of root, stem, leaf, and flower transcriptome data.
169

170 # 3. Results

171 ## 3.1. Near telomere-to-telomere level genome assembly of marigold

172 A total of 49.7 Gb (64 ×) PacBio HiFi reads with a read N50 length of 18.3 Kb (**Supplementary**
173 **Table S1**) were assembled into contigs using Hifiasm (Cheng et al., 2021). The primary assembly
174 includes 381 contigs with a total length of 1,011 Mb, a contig N50 size of 36.5 Mb, and a contig

175 N90 size of 8.8 Mb (**Supplementary Table S2**). This assembly is obviously larger than the
176 estimated genome size (~780 Mb) that evaluated based on the distribution of k-mer frequencies
177 (**Supplementary Figure S1**), which may result from genome heterozygosity. After purging the
178 heterozygous contigs from the primary assembly using Purge_dups (Guan et al., 2020), we obtained
179 a genome assembly with a total length of 778.9 Mb (**Table 1**), which is comparable to the estimated
180 genome size. The genome assembly comprises 47 contigs with contig N50 and N90 size of 38.0 and
181 11.0 Mb, respectively (**Table 1 and Supplementary Table S2**), revealing a BUSCO complete rate
182 of 98.3% (**Table 1**). In addition, the contig correctness was evaluated by using yak, which revealed
183 an average base accuracy of 99.9998% (QV57), suggesting the high accuracy of our genome
184 assembly. Compared to recently published genome assembly of Marigold by Xin et al. (defined as
185 assembly v1) (Xin et al., 2023), the huge improvement of our genome assembly (defined as
186 assembly v2) was highlighted by the total contig number (~73-fold decrease), contig N50 size (27-
187 folds increase), and genome size (71.7 Mb increase) (**Table 1**). Based on the Hi-C data
188 (**Supplementary Table S1 and S3**), we successfully anchored 99.7% (776.8 Mb) of contig
189 sequences onto 12 pseudo-chromosomes (**Table 1, Figure 1a and b**), which was obviously higher
190 than that of v1 assembly (87.6%) (**Table 1**). Finally, we obtained a near telomere-to-telomere level
191 reference genome for marigold, with scaffold N50 size of 69.8 Mb and N90 size of 40.8 Mb.
192
193 We identified a total of 590.4 Mb (75.8%) of TEs in v2 assembly (**Table 1 and Supplementary**
194 **Figure S2**), including 84.1 Mb (14.2%) of structural intact TEs. The proportion of TEs in v2
195 assembly falls in the range of that for the closely related species (73.4% - 84.7%) (Badouin et al.,
196 2017; Fan et al., 2022; Liu et al., 2020; Xu et al., 2021), and showed a 18.1% increase compared to

197 v1 assembly (57.7%) (**Table 1 and Supplementary Table S4**). In addition, we identified a total of
198 242.8 Mb (31.2% of the genome) TRs in marigold (**Table 1**). Then, we compared the TEs and TRs
199 in v2 assembly, and found that 97.5% of TRs were overlapped with TEs (**Supplementary Figure**
200 **S3**), which means that most of the TRs were formed from tandem clusters of the same type of TEs.
201
202 We predicted 42,529 protein-coding genes in v2 assembly, similar to that of *Mikania micrantha*
203 (46,351), *Helianthus annuus* (57,126), and *Stevia rebaudiana* (44,143), and obviously fewer than
204 that of *Smallanthus sonchifolius* (89,959) (Badouin et al., 2017; Fan et al., 2022; Liu et al., 2020;
205 Xu et al., 2021), which experienced a recent whole genome duplication (Fan et al., 2022). The
206 genome has a mean coding sequence length of 1,158 bp and exon number of 5.1, which was similar
207 to that of the related species (**Supplementary Table S5**). Functional annotation showed that 86.7%
208 of genes could be annotated by one of the NCBI-NR, KEGG, InterPro or GO databases
209 (**Supplementary Table S6**). Compared to v1 assembly, v2 assembly identified more genes (8,705
210 genes), which resulted in a 9.0% increase in BUSCO complete rate (**Table 1 and Supplementary**
211 **Table S5**), indicating the higher completeness of the v2 gene set. In addition, we identified 1,098
212 tRNA genes, 2,126 rRNA genes, and 929 other non-coding RNA genes in v2 assembly
213 (**Supplementary Table S7**).
214

215 **3.2. Comprehensive assembly of repetitive regions and correction of large-scale**
216 **assembly errors**

217 In v2 assembly, 79.2% of the assembled chromosome-ends have telomeric repeats (**Table 1**),
218 including 7 pseudo-chromosomes have telomeric repeats at both ends and 5 pseudo-chromosomes

219 have telomeric repeats at a single end (**Figure 1b and Supplementary Figure S4**). Compared to
220 v1 assembly, the v2 assembly reconstructed obviously more telomeres (79.2% vs 8.3%) (**Table 1**),
221 indicating the improvement in assembly of the ends for pseudo-chromosomes.
222
223 Compared to v1 assembly, the v2 assembly has an obvious increase (155.6 Mb) in the total length
224 of the 12 pseudo-chromosomes, which caused by both the higher total contig size (778.9 Mb vs
225 707.2 Mb) and the higher anchor rate (99.7% vs 87.6%) in v2 assembly (**Table 1**). Comparisons of
226 the pseudo-chromosomes between the assembly v2 and v1 revealed an obviously one-to-one match
227 relationship for all 12 pseudo-chromosomes (**Figure 2**). Compared to v2 assembly, large blocks
228 were missed in v1 assembly for many pseudo-chromosomes, including Chr02 (33.1 Mb), Chr03
229 (27.5 Mb), Chr04 (9.1 Mb), Chr05 (32.4 Mb), Chr06 (10.0 Mb), Chr08 (20.1 Mb), and Chr10 (25.5
230 Mb) (**Figure 2 and Supplementary Table S8**). After annotating of these large blocks in v2
231 assembly, we found that most of these sequences were identified as TRs, which were barely
232 assembled in v1 assembly (**Figure 2**).
233
234 Moreover, the pseudo-chromosomes with large blocks that showed inconsistent ordering or
235 orientation in v1 assembly were analyzed. For Chr03, the whole pseudo-chromosome in v1
236 assembly was split into two large blocks by a huge gap according to the alignment result, and the
237 ordering of the two blocks was inconsistent with that in v2 assembly (**Figure 2**). Similar results
238 were observed for Chr04, Chr05, and Chr08 (**Figure 2**). Because the huge gap will greatly reduce
239 the Hi-C linkage signal between the two blocks (Zhang et al., 2019), it is easy to result in wrong
240 connections among different chromosome-ends in v1 assembly. In addition, no telomeric repeats

241 was identified from the ends of these pseudo-chromosomes in v1 assembly, but identified from the
242 single or both ends in v2 assembly (**Figure 1b and Supplementary Figure S4**), which suggested
243 the correctness of these pseudo-chromosomes in v2 assembly. For Chr01, Chr06 and Chr11, the
244 chromosomes in v1 assembly were split into many large blocks, which were inconsistent with that
245 in v2 assembly, though the length of these chromosomes are similar (**Figure 2**). Since v1 assembly
246 has telomeric repeats located in the sub-terminal of the three pseudo-chromosomes instead of the
247 chromosome-ends, while v2 assembly has telomeric repeats at single or both ends (**Figure 1b and**
248 **Supplementary Figure S4**), we inferred that the v2 assembly has corrected the order and
249 orientation mistakes in v1 assembly. In total, the v2 assembly corrected 7 pseudo-chromosomes that
250 contained wrong ordering or orientation blocks in v1 assembly.
251

252 **3.3. High proportion of centromeric 170-bp unit reveals the existence of**
253 **holocentromeres**

254 The Hi-C heatmap showed many large blocks in genome assembly with no or weak Hi-C linkage
255 signals (**Figure 1a**), revealing a high proportion of high similar repeating sequences existed in the
256 genome. We identified a total of 242.8 Mb (31.2% of the genome) TRs, and the distribution of the
257 TRs was highly consistent with the location of these large blocks in Hi-C heatmap (**Supplementary**
258 **Figure S5**). Noticeably, the proportion of TRs in marigold genome was obviously higher than that
259 of *Smallanthus sonchifolius* (101.3 Mb, 3.7% of the genome), *Mikania micrantha* (108.6 Mb, 6.1%
260 of the genome), *Helianthus annuus* (163.1 Mb, 5.4% of the genome), and *Stevia rebaudiana* (72.8
261 Mb, 5.1% of the genome). In addition, the TRs and TEs were highly overlapped in marigold genome
262 (**Supplementary Figure S3**), and more than 97.5% of TRs were classified as TEs.

263

264 Previous studies have reported that centromere consisted of highly repetitive DNA sequences
265 with monomers of length 100 - 200 bp (Melters et al., 2013; Naish et al., 2021; Shi et al., 2023;
266 Sundararajan and Straight, 2022). Here, we found that 170-bp repeats were the most abundant unit
267 in the genome, which had 682,138 repetitions accounted for 14.8% of the total genome sequences,
268 followed by 167-bp (2.7%), 178-bp (1.2%) and 169-bp (0.9%) repeats (**Supplementary Table S9**).
269 These four repeat units localized on all 12 pseudo-chromosomes, with the total length ranged from
270 4.0 Kb (Chr09) to 29.0 Mb (Chr02) (**Supplementary Table S10**). In detail, the 170-bp was the
271 main repeat unit in Chr01, Chr02, Chr03, Chr05, Chr06, and Chr07, and the 169-bp was the main
272 repeat unit in Chr11 (**Figure 3a and Supplementary Figure S6**). In addition, the 167-bp and 170-
273 bp repeats were the main units for Chr04, Chr08, and Chr10, the 167-bp and 178-bp repeats were
274 the main units for Chr09, and 169-bp and 170-bp repeats were the main units for Chr12 (**Figure 3a**
275 **and Supplementary Figure S6**). Noticeably, these repeat units distributed across the whole
276 chromosome for all 12 pseudo-chromosomes, but not constrained in a single region (**Figure 3b**).
277 Besides, these repeat regions have obviously lower gene density than that of other parts on all
278 pseudo-chromosomes (**Figure 3b**). Moreover, more than 80.2% of sequences within these repeat
279 regions were annotated as retrotransposons (**Figure 3b**), which was consistent with that centromeres
280 are composed of long arrays of tandem repeats and/or retrotransposons in many species (Henikoff
281 et al., 2001; Zhang et al., 2014). Based on these results, we inferred that these repeat regions in
282 marigold genome were centromeric regions, and the chromosomes of marigold should be classified
283 as holocentric chromosomes, similar to *Holhymenia histrio*, *Rhynchospora pubera*, and *Vitis*
284 *vinifera* (Bardella et al., 2020; Hofstatter et al., 2022; Marques et al., 2015; Shi et al., 2023).

285

286 **3.4. Differential expression patterns of the four *PSY* genes in different tissues of**

287 **marigold**

288 In marigold, lutein accounts for up to 90% of the total carotenoid contents in the flower petals, and

289 all genes involved in lutein biosynthesis were identified by Xin *et al*. based on a whole genome

290 sequence (Xin et al., 2023). Phytoene synthase (PSY) catalyzes the first and rate-limiting step in

291 lutein biosynthesis (**Figure 4a**), and four *PSY* genes were identified from marigold genome (Xin et

292 al., 2023). However, the structure and the phylogenetic relationship of the four genes in marigold

293 were not investigated. Based on our high-quality genome assembly, we identified four *PSY* genes

294 (*PSYa*, *PSYb*, *PSYc*, and *PSYd*) (**Supplementary Figure S7**), consistent with the previous study

295 (Xin et al., 2023). The four *PSY* genes have a length of 2,841 bp, 2,760bp, 3,235 bp, and 1,917 bp,

296 respectively, all of which contain 6 exons and 5 introns (**Supplementary Figure S7**). The predicted

297 protein sequences have 425, 398, 427, and 375 amino acids for *PSYa*, *PSYb*, *PSYc*, and *PSYd* genes,

298 respectively. To investigate the phylogenetic relationship of these four *PSY* genes, 28 PSY protein

299 sequences from four Solanales species and five Asterales species were used (**Figure 4b**). The

300 phylogenetic tree showed that *PSYa* and *PSYc* were tightly grouped together, which was then

301 clustered with *PSYb* (**Figure 4b**). In addition, *PSYd* was clustered into another group, and showed

302 a high divergence to other three *PSY* genes (**Figure 4b**), which was consistent with that *PSYd*

303 showed a low amino acid sequence identity to *PSYa*, *PSYb* and *PSYc* (60.9 - 62.2%).

304

305 Based on the recently published marigold genome, Xin *et al*. investigated the expression patterns

306 of lutein biosynthesis genes at four flowering development stages in marigold flowers (Xin et al.,

307 2023), but the differences among different tissues were not studied. In this study, we compared the
308 expression levels of these genes among different tissues, including roots, stems, leaves, and flowers.
309 For *PSY* genes, *PSYa* only expressed in flowers and showed a high expression level, *PSYb* showed
310 moderate expression levels in flowers and leaves, and low level in stems, and *PSYc* exhibited
311 moderate expression levels in leaves, and low level in stems and roots (**Figure 4c**). In addition,
312 *PSYd* only expressed in roots with very low expression level (**Figure 4c**), which was similar to its
313 homolog *SlPSY3* in tomato that showed low but detectable transcript levels in roots (Fantini et al.,
314 2013). These results suggested that the four *PSY* genes have diversified and possibly executed
315 different functions in roots, stems, leaves, and flowers. In addition, the expression levels of the other
316 genes were all obviously higher in flowers than those in leaves, stems, and roots (**Figure 4c**), which
317 was consistent with that the flowers accumulated with the highest level of lutein in marigold (Moehs
318 et al., 2001).
319

320 **4. Discussions**

321 In this study, we reported a near telomere-to-telomere level genome assembly of marigold based on
322 highly accurate HiFi long reads and Hi-C sequencing data. Compared to the previously reported
323 marigold genome, the current assembly had obviously higher contiguity and higher completeness
324 of gene set. Besides, the current assembly has much fewer assembly errors. Although all the
325 assembled pseudo-chromosomes contain less than 5 contigs in this study, a few gaps still exist in
326 the genome assembly. In addition, we have assembled telomere sequences for 19 chromosome-ends,
327 but 5 chromosome-ends still have no assembled telomere sequences. To generate a real telomere-

328 to-telomere genome assembly, further investigations based on ultra-long sequencing and other
329 technologies will be necessary.
330
331 Based on the high-quality genome assembly, we found that the 170-bp repeats are the most
332 abundant centromeric unit, which followed by 167-bp, 178-bp and 169-bp repeats. All centromeric
333 units are distributed along the whole chromosomes for all 12 centromeres, indicating the existence
334 of the holocentromeres in marigold. Here, the centromeres were identified mainly based on the
335 analysis of the tandem repeat units. To ascertain the centromere type of marigold, more studies are
336 needed to investigate the organization of spindle fibres at metaphase, since the previous studies
337 reported that the spindle fibres are attached along almost the entire poleward surface of the
338 chromatids in holocentromeres (Schubert et al., 2020). In addition, the chromosomal distribution of
339 the centromere-specific histone H3 (CENH3) protein and the cell cycle-dependent pericentromeric
340 phosphorylation of histone H3 and H2A need further investigations (Baez et al., 2020).
341
342 The reference genome sequence of marigold greatly facilitates lutein biosynthesis studies. Xin *et*
343 *al*. have identified all the genes involved in lutein biosynthesis based on a low-continuity reference
344 genome of marigold, and investigated the expression patterns of these genes at four flowering
345 development stages in flowers (Xin et al., 2023). However, in marigold, the relationship of the four
346 *PSY* genes, and the expression patterns of these genes among different tissues were not studied. In
347 this study, we analyzed the structure and phylogenetic relationship of the four *PSY* genes, revealing
348 that *PSYa* and *PSYc* were tightly grouped together, then clustered with *PSYb*, and *PSYd* formed a
349 separate group that with a high divergence to other three *PSY* genes. In addition, our results revealed

that the four *PSY* genes have diversified and possibly executed different functions in roots, stems, leaves, and flowers. To verify the functions of these *PSY* genes, more investigations based on the molecular genetic methods, such as gene editing technologies, are needed. Our near telomere-to-telomere level genome assembly and comprehensive gene annotation will greatly facilitate the breeding of marigold and researches aimed at improving lutein production.

## Data Availability Statement

All raw sequencing data generated during the current study have been deposited at DDBJ/ENA/GenBank under project accession PRJNA975275. Genomic sequence reads have been deposited in the SRA database with accession SRR24689467 for PacBio sequencing. Hi-C sequencing reads have been deposited in the SRA database with accession SRR24718810. Full-length transcript sequence reads have been deposited in the SRA database with accession SRR24733750. Illumina transcript sequences for roots, stems, leaves, and flowers have been deposited in SRA database with accession SRR24733730-SRR24733749. The genome assembly and annotation have deposited at NCBI with accession JAUHHV000000000, and also available at China National Genomics Data Center (https://ngdc.cncb.ac.cn) with the Project ID PRJCA017901 and Zenodo: https://www.zenodo.org/record/8107331.

## CRediT authorship contribution statement

**Fan Jiang**: Investigation, Methodology, Formal analysis, Writing - original draft. **Lihua Yuan**: Investigation, Methodology, Formal analysis. **Sen Wang**: Formal analysis, Validation. **Hengchao**

**Wang**: Formal analysis. **Dong Xu**: Formal analysis. **Anqi Wang**: Validation, Software. **Wei Fan**: Supervision, Writing - review & editing, Funding acquisition.

## Declaration of competing Interest

The authors declare no competing interest.

## Acknowledgment

The work was funded by the Agricultural Science and Technology Innovation Program and the Elite Young Scientists Program of CAAS, the fund of Key Laboratory of Shenzhen (ZDSYS20141118170111640), and the National Natural Science Foundation of China (32172430).

398

399

400 ## References


401 Badouin H, Gouzy J, Grassa CJ, Murat F, Staton SE, Cottret L, Lelandais-Briere C, Owens GL,
402 Carrere S, Mayjonade B, Legrand L, Gill N, Kane NC, Bowers JE, Hubner S, Bellec A,
403 Berard A, Berges H, Blanchet N, Boniface MC, Brunel D, Catrice O, Chaidir N, Claudel
404 C, Donnadieu C, Faraut T, Fievet G, Helmstetter N, King M, Knapp SJ, Lai Z, Le Paslier
405 MC, Lippi Y, Lorenzon L, Mandel JR, Marage G, Marchand G, Marquand E, Bret-Mestries
406 E, Morien E, Nambeesan S, Nguyen T, Pegot-Espagnet P, Pouilly N, Raftis F, Sallet E,
407 Schiex T, Thomas J, Vandecasteele C, Vares D, Vear F, Vautrin S, Crespi M, Mangin B,
408 Burke JM, Salse J, Munos S, Vincourt P, Rieseberg LH, and Langlade NB, 2017. The
409 sunflower genome provides insights into oil metabolism, flowering and Asterid evolution.
410 Nature. 546 (7656), 148. https://doi.org/10.1038/nature22380.

411 Baez M, Kuo YT, Dias Y, Souza T, Boudichevskaia A, Fuchs J, Schubert V, Vanzela ALL, Pedrosa-
412 Harand A, and Houben A, 2020. Analysis of the small chromosomal Prionium serratum
413 (Cyperid) demonstrates the importance of reliable methods to differentiate between mono-
414 and holocentricity. Chromosoma. 129 (3-4), 285-97. https://doi.org/10.1007/s00412-020-
415 00745-6.

416 Bao W, Kojima KK, and Kohany O, 2015. Repbase Update, a database of repetitive elements in
417 eukaryotic genomes. Mob. DNA. 6, 11. https://doi.org/10.1186/s13100-015-0041-9.

418 Bardella VB, Milani D, and Cabral-de-Mello DC, 2020. Analysis of Holhymenia histrio genome
419 provides insight into the satDNA evolution in an insect with holocentric chromosomes.

420 Chromosome Research. 28 (3), 369-80. https://doi.org/10.1007/s10577-020-09642-1.

421 Benson G, 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res.
422 27 (2), 573-80. https://doi.org/10.1093/nar/27.2.573.

423 Blum M, Chang HY, Chuguransky S, Grego T, Kandasaamy S, Mitchell A, Nuka G, Paysan-Lafosse
424 T, Qureshi M, Raj S, Richardson L, Salazar GA, Williams L, Bork P, Bridge A, Gough J,
425 Haft DH, Letunic I, Marchler-Bauer A, Mi H, Natale DA, Necci M, Orengo CA,
426 Pandurangan AP, Rivoire C, Sigrist CJA, Sillitoe I, Thanki N, Thomas PD, Tosatto SCE,
427 Wu CH, Bateman A, and Finn RD, 2021. The InterPro protein families and domains
428 database: 20 years on. Nucleic Acids Res. 49 (D1), D344-D54.
429 https://doi.org/10.1093/nar/gkaa977.

430 Buchfink B, Reuter K, and Drost HG, 2021. Sensitive protein alignments at tree-of-life scale using
431 DIAMOND. Nat. Methods. 18 (4), 366. https://doi.org/10.1038/s41592-021-01101-x.

432 Burlec AF, Pecio L, Kozachok S, Mircea C, Corciova A, Verestiuc L, Cioanca O, Oleszek W, and
433 Hancianu M, 2021. Phytochemical profile, antioxidant activity, and cytotoxicity
434 assessment of Tagetes erecta L. flowers. Molecules. 26 (5), 1201.
435 https://doi.org/10.3390/molecules26051201.

436 Buscemi S, Corleo D, Di Pace F, Petroni ML, Satriano A, and Marchesini G, 2018. The effect
437 oflutein on eye and extra-eye health. Nutrients. 10 (9), 1321.
438 https://doi.org/10.3390/nu10091321.

439 Chan PP, Lin BY, Mak AJ, and Lowe TM, 2021. tRNAscan-SE 2.0: improved detection and
440 functional classification of transfer RNA genes. Nucleic Acids Res. 49 (16), 9077-96.
441 https://doi.org/10.1093/nar/gkab688.

442 Cheng HY, Concepcion GT, Feng XW, Zhang HW, and Li H, 2021. Haplotype-resolved de novo
443 assembly using phased assembly graphs with hifiasm. Nat. Methods. 18 (2), 170.
444 https://doi.org/10.1038/s41592-020-01056-5.
445 Chung RWS, Leanderson P, Lundberg AK, and Jonasson L, 2017. Lutein exerts anti-inflammatory
446 effects in patients with coronary artery disease. Atherosclerosis. 262, 87-93.
447 https://doi.org/10.1016/j.atherosclerosis.2017.05.008.
448 da Cruz MHP, Domingues DS, Saito PTM, Paschoal AR, and Bugatti PH, 2021. TERL:
449 classification of transposable elements by convolutional neural networks. Brief. Bioinform.
450 22 (3), 1-15. https://doi.org/10.1093/bib/bbaa185.
451 Emms DM and Kelly S, 2019. OrthoFinder: phylogenetic orthology inference for comparative
452 genomics. Genome Biol. 20 (1), 238. https://doi.org/10.1186/s13059-019-1832-y.
453 Fan W, Wang S, Wang HC, Wang AQ, Jiang F, Liu HW, Zhao HB, Xu D, and Zhang Y, 2022. The
454 genomes of chicory, endive, great burdock and yacon provide insights into Asteraceae
455 palaeo-polyploidization history and plant inulin production. Molecular Ecology Resources.
456 22 (8), 3124-40. https://doi.org/10.1111/1755-0998.13675.
457 Fantini E, Falcone G, Frusciante S, Giliberto L, and Giuliano G, 2013. Dissection of Tomato
458 Lycopene Biosynthesis through Virus-Induced Gene Silencing. Plant Physiology. 163 (2),
459 986-98. https://doi.org/10.1104/pp.113.224733.
460 Feng G, Huang S, Liu Y-R, Xiao F, Liu J, Zhang Z, Chen Q, Mao Y, Cao XR, Wang Y, Chen D,
461 Zhou Y, Yu F-M, Liu G, Liu Y, and Niu X, 2018. The transcriptome analyses of Tagetes
462 erecta provides novel insights into secondary metabolite biosynthesis during flower
463 development. Gene. 660, 18-27.

464 Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, and Smit AF, 2020.
465 RepeatModeler2 for automated genomic discovery of transposable element families. Proc.
466 Natl. Acad. Sci. U.S.A. 117 (17), 9451-57. https://doi.org/10.1073/pnas.1921046117.
467 Gansukh E, Mya KK, Jung M, Keum YS, Kim DH, and Saini RK, 2019. Lutein derived from
468 marigold (Tagetes erecta) petals triggers ROS generation and activates Bax and caspase-3
469 mediated apoptosis of human cervical carcinoma (HeLa) cells. Food Chem. Toxicol. 127,
470 11-18. https://doi.org/10.1016/j.fct.2019.02.037.
471 Giuliano G, 2014. Plant carotenoids: genomics meets multi-gene engineering. Curr Opin Plant Biol.
472 19, 111-7. https://doi.org/10.1016/j.pbi.2014.05.006.
473 Guan D, McCarthy SA, Wood J, Howe K, Wang Y, and Durbin R, 2020. Identifying and removing
474 haplotypic duplication in primary genome assemblies. Bioinformatics. 36 (9), 2896-98.
475 https://doi.org/10.1093/bioinformatics/btaa025.
476 Henikoff S, Ahmad K, and Malik HS, 2001. The centromere paradox: stable inheritance with rapidly
477 evolving DNA. Science. 293 (5532), 1098-102. https://doi.org/10.1126/science.1062939.
478 Hofstatter PG, Thangavel G, Lux T, Neumann P, Vondrak T, Novak P, Zhang M, Costa L, Castellani
479 M, Scott A, Toegelova H, Fuchs J, Mata-Sucre Y, Dias Y, Vanzela ALL, Huettel B, Almeida
480 CCS, Simkova H, Souza G, Pedrosa-Harand A, Macas J, Mayer KFX, Houben A, and
481 Marques A, 2022. Repeat-based holocentromeres influence genome architecture and
482 karyotype evolution. Cell. 185 (17), 3153-68 e18.
483 https://doi.org/10.1016/j.cell.2022.06.045.
484 Kim D, Paggi JM, Park C, Bennett C, and Salzberg SL, 2019. Graph-based genome alignment and
485 genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 37 (8), 907-15.

486 https://doi.org/10.1038/s41587-019-0201-4.

487 Kim J, Smith JJ, Tian L, and Dellapenna D, 2009. The evolution and function of carotenoid
488 hydroxylases in Arabidopsis. Plant Cell Physiol. 50 (3), 463-79.
489 https://doi.org/10.1093/pcp/pcp005.

490 Langmead B and Salzberg SL, 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods. 9
491 (4), 357-9. https://doi.org/10.1038/nmeth.1923.

492 Li LH, Lee JC, Leung HH, Lam WC, Fu Z, and Lo ACY, 2020. Lutein supplementation for eye
493 diseases. Nutrients. 12 (6), 1721. https://doi.org/10.3390/nu12061721.

494 Lin JH, Lee DJ, and Chang JS, 2015. Lutein production from biomass: marigold flowers versus
495 microalgae. Bioresour. Technol. 184, 421-28.
496 https://doi.org/10.1016/j.biortech.2014.09.099.

497 Liu B, Shi Y, Yuan J, Hu X, Zhang H, Li N, Li Z, Chen Y, Mu D, and Fan W, 2013. Estimation of
498 genomic characteristics by analyzing k-mer frequency in de novo genome projects. arXiv:
499 Genomics.

500 Liu B, Yan J, Li WH, Yin LJ, Li P, Yu HX, Xing LS, Cai ML, Wang HC, Zhao MX, Zheng J, Sun
501 F, Wang ZZ, Jiang ZY, Ou QJ, Li SB, Qu L, Zhang QL, Zheng YP, Qiao X, Xi Y, Zhang Y,
502 Jiang F, Huang C, Liu CH, Ren YW, Wang S, Liu HW, Guo JY, Wang HH, Dong H, Peng
503 CL, Qian WQ, Fan W, and Wan FH, 2020. Mikania micrantha genome provides insights
504 into the molecular mechanism of rapid growth. Nat. Commun. 11 (1), 340.
505 https://doi.org/ARTN 340

506 10.1038/s41467-019-13926-4.

507 Maass D, Arango J, Wust F, Beyer P, and Welsch R, 2009. Carotenoid crystal formation in

508 Arabidopsis and carrot roots caused by increased phytoene synthase protein levels. PLoS
509 One. 4 (7), e6373. https://doi.org/10.1371/journal.pone.0006373.
510 Marques A, Ribeiro T, Neumann P, Macas J, Novak P, Schubert V, Pellino M, Fuchs J, Ma W,
511 Kuhlmann M, Brandt R, Vanzela ALL, Beseda T, Simkova H, Pedrosa-Harand A, and
512 Houben A, 2015. Holocentromeres in Rhynchospora are associated with genome-wide
513 centromere-specific repeat arrays interspersed among euchromatin. Proc. Natl. Acad. Sci.
514 USA. 112 (44), 13633-38. https://doi.org/10.1073/pnas.1512255112.
515 Melters DP, Bradnam KR, Young HA, Telis N, May MR, Ruby JG, Sebra R, Peluso P, Eid J, Rank
516 D, Garcia JF, DeRisi JL, Smith T, Tobias C, Ross-Ibarra J, Korf I, and Chan SWL, 2013.
517 Comparative analysis of tandem repeats from hundreds of species reveals unique insights
518 into centromere evolution. Genome Biol. 14 (1), R10. https://doi.org/ARTN R10
519 10.1186/gb-2013-14-1-r10.
520 Moehs CP, Tian L, Osteryoung KW, and DellaPenna D, 2001. Analysis of carotenoid biosynthetic
521 gene expression during marigold petal development. Plant Molecular Biology. 45 (3), 281-
522 93. https://doi.org/Doi 10.1023/A:1006417009203.
523 Moliner C, Barros L, Dias MI, Lopez V, Langa E, Ferreira ICFR, and Gomez-Rincon C, 2018.
524 Edible flowers of Tagetes erecta L. as functional ingredients: phenolic composition,
525 antioxidant and protective effects on Caenorhabditis elegans. Nutrients. 10 (12), 2002.
526 https://doi.org/ARTN 2002
527 10.3390/nu10122002.
528 Naish M, Alonge M, Wlodzimierz P, Tock AJ, Abramson BW, Schmucker A, Mandakova T, Jamge
529 B, Lambing C, Kuo P, Yelina N, Hartwick N, Colt K, Smith LM, Ton J, Kakutani T,

530 Martienssen RA, Schneeberger K, Lysak MA, Berger F, Bousios A, Michael TP, Schatz
531 MC, and Henderson IR, 2021. The genetic and epigenetic landscape of the Arabidopsis
532 centromeres. Science. 374 (6569), eabi7489. https://doi.org/10.1126/science.abi7489.
533 Nawrocki EP and Eddy SR, 2013. Infernal 1.1: 100-fold faster RNA homology searches.
534 Bioinformatics. 29 (22), 2933-5. https://doi.org/10.1093/bioinformatics/btt509.
535 Nisar N, Li L, Lu S, Khin NC, and Pogson BJ, 2015. Carotenoid metabolism in plants. Mol Plant.
536 8 (1), 68-82. https://doi.org/10.1016/j.molp.2014.12.007.
537 Nwachukwu ID, Udenigwe CC, and Aluko RE, 2016. Lutein and zeaxanthin: Production technology,
538 bioavailability, mechanisms of action, visual function, and health claim status. Trends in
539 Food Science & Technology. 49, 74-84. https://doi.org/10.1016/j.tifs.2015.12.005.
540 Ou S, Su W, Liao Y, Chougule K, Agda JRA, Hellinga AJ, Lugo CSB, Elliott TA, Ware D, Peterson
541 T, Jiang N, Hirsch CN, and Hufford MB, 2019. Benchmarking transposable element
542 annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biol.
543 20 (1), 275. https://doi.org/10.1186/s13059-019-1905-y.
544 Park YJ, Park SY, Valan Arasu M, Al-Dhabi NA, Ahn HG, Kim JK, and Park SU, 2017.
545 Accumulation of carotenoids and metabolic profiling in different cultivars of Tagetes
546 flowers. Molecules. 22 (2), 313. https://doi.org/10.3390/molecules22020313.
547 Price MN, Dehal PS, and Arkin AP, 2010. FastTree 2--approximately maximum-likelihood trees for
548 large alignments. PLoS One. 5 (3), e9490. https://doi.org/10.1371/journal.pone.0009490.
549 Rodriguez-Villalon A, Gas E, and Rodriguez-Concepcion M, 2009. Phytoene synthase activity
550 controls the biosynthesis of carotenoids and the supply of their metabolic precursors in
551 dark-grown Arabidopsis seedlings. Plant Journal. 60 (3), 424-35.

552 https://doi.org/10.1111/j.1365-313X.2009.03966.x.

553 Schubert V, Neumann P, Marques A, Heckmann S, Macas J, Pedrosa-Harand A, Schubert I, Jang
554 TS, and Houben A, 2020. Super-resolution microscopy reveals diversity of plant
555 centromere architecture. International Journal of Molecular Sciences. 21 (10), 3488.
556 https://doi.org/ARTN 3488
557 10.3390/ijms21103488.

558 Servant N, Varoquaux N, Lajoie BR, Viara E, Chen CJ, Vert JP, Heard E, Dekker J, and Barillot E,
559 2015. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. Genome Biol.
560 16, 259. https://doi.org/10.1186/s13059-015-0831-x.

561 Shi X, Cao S, Wang X, Huang S, Wang Y, Liu Z, Liu W, Leng X, Peng Y, Wang N, Wang Y, Ma Z,
562 Xu X, Zhang F, Xue H, Zhong H, Wang Y, Zhang K, Velt A, Avia K, Holtgrawe D, Grimplet
563 J, Matus JT, Ware D, Wu X, Wang H, Liu C, Fang Y, Rustenholz C, Cheng Z, Xiao H, and
564 Zhou Y, 2023. The complete reference genome for grapevine (Vitis vinifera L.) genetics
565 and breeding. Hortic Res. 10 (5), uhad061. https://doi.org/10.1093/hr/uhad061.

566 Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, and Zdobnov EM, 2015. BUSCO:
567 assessing genome assembly and annotation completeness with single-copy orthologs.
568 Bioinformatics. 31 (19), 3210-2. https://doi.org/10.1093/bioinformatics/btv351.

569 Slater GS and Birney E, 2005. Automated generation of heuristics for biological sequence
570 comparison. BMC Bioinformatics. 6, 31. https://doi.org/10.1186/1471-2105-6-31.

571 Smit A, Hubley R, and Green P, RepeatMasker Open-4.0. 2015. http://repeatmasker.org/. Accessed:
572 10 November, 2020.,

573 Stanke M, Schoffmann O, Morgenstern B, and Waack S, 2006. Gene prediction in eukaryotes with

574 a generalized hidden Markov model that uses hints from external sources. BMC
575 Bioinformatics. 7 (1), 62. https://doi.org/10.1186/1471-2105-7-62.
576 Sun TH, Yuan H, Cao HB, Yazdani M, Tadmor Y, and Li L, 2018. Carotenoid Metabolism in Plants:
577 The Role of Plastids. Molecular Plant. 11 (1), 58-74.
578 https://doi.org/10.1016/j.molp.2017.09.010.
579 Sundararajan K and Straight AF, 2022. Centromere Identity and the Regulation of Chromosome
580 Segregation. Front Cell Dev Biol. 10, 914249. https://doi.org/10.3389/fcell.2022.914249.
581 Tanaka Y, Sasaki N, and Ohmiya A, 2008. Biosynthesis of plant pigments: anthocyanins, betalains
582 and carotenoids. Plant J. 54 (4), 733-49. https://doi.org/10.1111/j.1365-313X.2008.03447.x.
583 Thorvaldsdottir H, Robinson JT, and Mesirov JP, 2013. Integrative Genomics Viewer (IGV): high-
584 performance genomics data visualization and exploration. Brief. Bioinform. 14 (2), 178-
585 92. https://doi.org/10.1093/bib/bbs017.
586 Wang S, Wang HC, Jiang F, Wang AQ, Liu HW, Zhao HB, Yang BY, Xu D, Zhang Y, and Fan W,
587 2022. EndHiC: assemble large contigs into chromosome-level scaffolds using the Hi-C
588 links from contig ends. BMC Bioinformatics. 23 (1), 528. https://doi.org/10.1186/s12859-
589 022-05087-x.
590 Xin H-b, Ji F, Wu J, Zhang S, Yi CJ, Zhao S, Cong R-C, Zhao L, Zhang H-l, and Zhang Z, 2023.
591 Chromosome-scale genome assembly of marigold (Tagetes erecta L.): an ornamental plant
592 and feedstock for industrial lutein production. Horticultural Plant Journal.
593 Xu XY, Yuan HY, Yu XQ, Huang SZ, Sun YM, Zhang T, Liu QQ, Tong HY, Zhang YX, Wang YJ,
594 Liu CX, Wu L, Hou ML, and Yang YH, 2021. The chromosome-level Stevia genome
595 provides insights into steviol glycoside biosynthesis. Horticulture Research. 8 (1), 129.

596 https://doi.org/ARTN 129
597 10.1038/s41438-021-00565-4.
598 Zhang C, Wei L, Wang W, Qi W, Cao Z, Li H, Bao M, and He Y, 2020a. Identification,
599 characterization and functional analysis of AGAMOUS subfamily genes associated with
600 floral organs and seed development in Marigold (Tagetes erecta). BMC Plant Biol. 20 (1),
601 439. https://doi.org/10.1186/s12870-020-02644-5.
602 Zhang H, Zhang S, Zhang H, Chen X, Liang F, Qin H, Zhang Y, Cong R, Xin H, and Zhang Z,
603 2020b. Carotenoid metabolite and transcriptome dynamics underlying flower color in
604 marigold (Tagetes erecta L.). Sci Rep. 10 (1), 16835. https://doi.org/10.1038/s41598-020-
605 73859-7.
606 Zhang H, Koblizkova A, Wang K, Gong Z, Oliveira L, Torres GA, Wu Y, Zhang W, Novak P, Buell
607 CR, Macas J, and Jiang J, 2014. Boom-Bust Turnovers of Megabase-Sized Centromeric
608 DNA in Solanum Species: Rapid Evolution of DNA Sequences Associated with
609 Centromeres. Plant Cell. 26 (4), 1436-47. https://doi.org/10.1105/tpc.114.123877.
610 Zhang X, Zhang S, Zhao Q, Ming R, and Tang H, 2019. Assembly of allele-aware, chromosomal-
611 scale autopolyploid genomes based on Hi-C data. Nat Plants. 5 (8), 833-45.
612 https://doi.org/10.1038/s41477-019-0487-8.
613
614
615
616
617
618
619

620
621
622
623 ## Table and legends

624 Table 1. Comparisons of the genomic assembly and annotation for marigold v2 and v1 assembly.

| Characteristics | v2 | v1 |
|---|---|---|
| **Genome assembly** | | |
| Contig number | 47 | 3,451 |
| Contig size (bp) | 778,905,021 | 707,210,696 |
| Contig N50 size (bp) | 37,953,657 | 1,391,023 |
| Scaffold number | 26 | 1,651 |
| Scaffold size (bp) | 778,926,021 | 709,010,696 |
| Scaffold N50 size (bp) | 69,673,756 | 54,151,473 |
| Pseudo-chromosomes length (bp) | 776,792,443 | 621,195,486 |
| Anchored to Pseudo-chromosomes (%) | 99.7 | 87.6 |
| Ratio of telomeres assembled | 19/24 | 2/24 |
| BUSCO complete rate of the genome (%) | 98.3 | 98.2 |
| **Genome annotation** | | |
| Transposable element length (bp) | 590,391,399 | 407,962,073 |
| % transposable elements | 75.8 | 57.5 |
| Tandem repeat length (bp) | 243,461,285 | 60,279,323 |
| % tandem repeats | 31.3 | 8.5 |
| Protein-coding gene number | 42,529 | 33,824 |
| Total CDS size (bp) | 49,266,489 | 38,796,777 |
| BUSCO complete rate of the genes (%) | 98.3 | 89.3 |

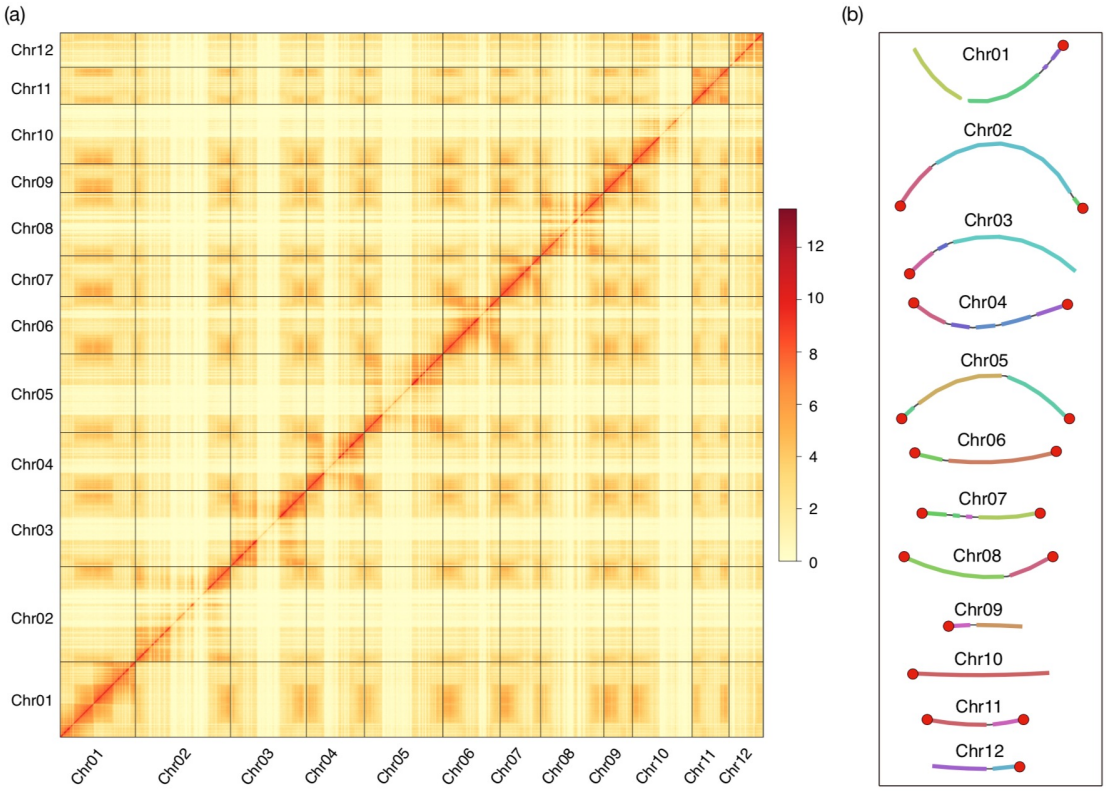
625
626
627
628
629

630

631

632

633 **Figure legends**




634

635 **Figure 1.** Assessment of the genome assembly for marigold. (a) Hi-C heatmap of the genome

636 assembly. We scanned the genome by 200-Kb non-overlapping window as a bin and calculated

637 valid interaction links of Hi-C data between any pair of bins, and color represents Log2(links

638 number); (b) View of the pseudo-chromosomes. The thick lines represent contigs, and the thin lines

639 represent the links between the two contigs. The chromosome ends assembled with telomere-

640 specific repeats (CCCTAAA) were highlighted with red solid circle.

641

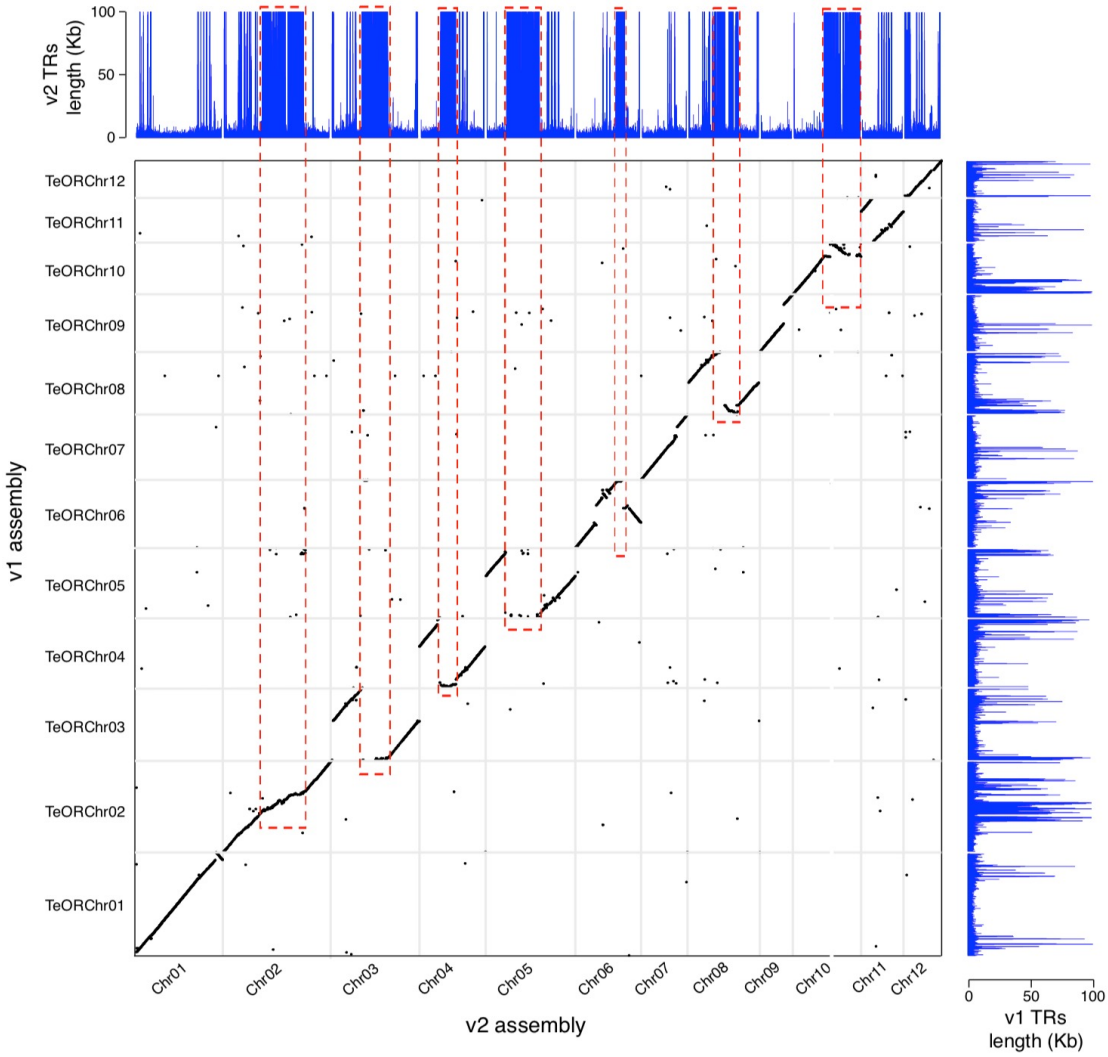



**Figure 2.** Comparisons between the genomes of marigold v2 and v1 assembly. The above and right parts showed the distributions of tandem repeats in v2 and v1 assembly, respectively. We scanned the genome by 100-Kb non-overlapping window as a bin to calculate the length of tandem repeats. The core part of the figure showed the pair-wise alignment of genome sequences between the v2 and v1 assembly. The corresponding segments for highly enriched tandem repeats in v2 assembly were highlighted with red box.

**Figure 3.** Characterization of holocentromeres in marigold genome. (a) Views of genes and four centromeric repeat units for Chr02 with IGV. We take Chr02 as an example to show the distributions of genes and different centromeric units. The head number showed the total length of the chromosome. (b) Distribution of genes, centromeric repeat units and the retrotransposons in centromeric regions for all 12 pseudo-chromosomes. The left columns showed the distribution of genes. The length represents Log2(length) in 100 Kb windows.

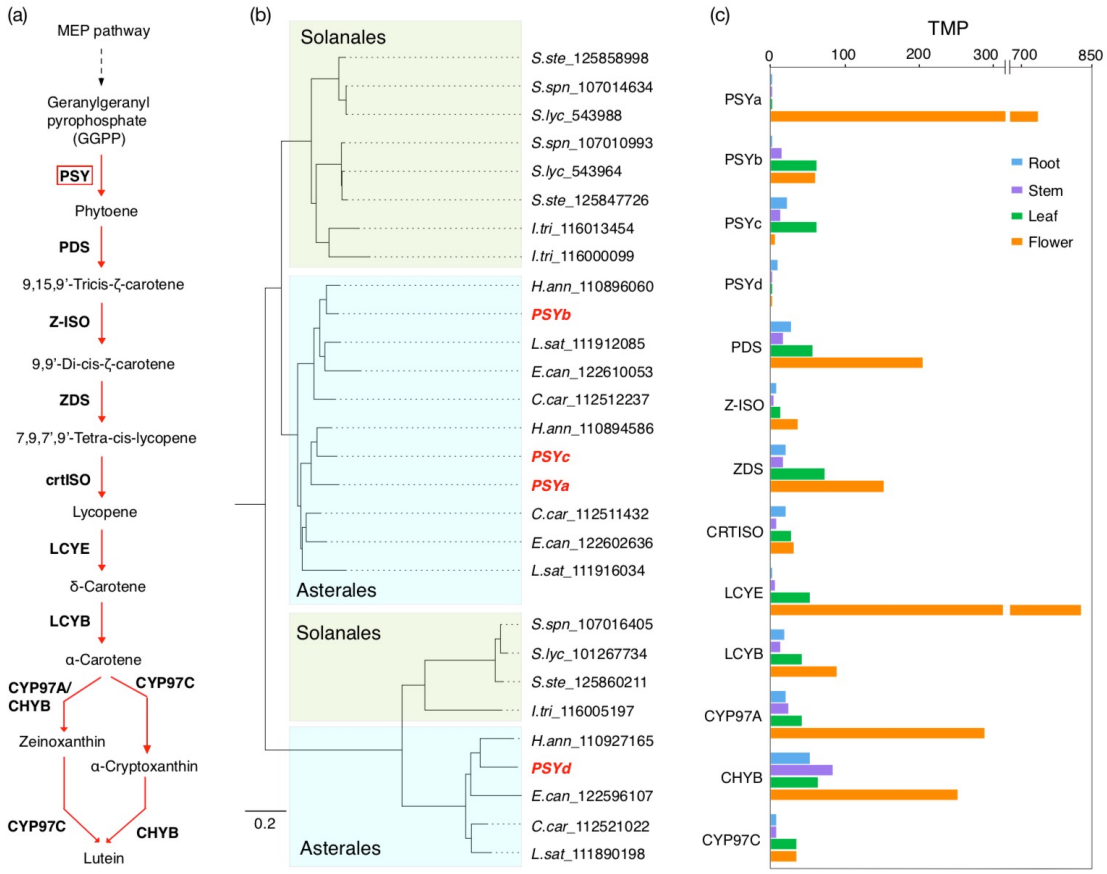



**Figure 4.** Analysis of lutein biosynthesis in marigold. (a) Primary steps for lutein biosynthesis pathway. (b) Phylogenetic tree for *PSY* genes from some Solanales and Asterales species. (c) Comparisons of the expression levels of all genes involved in lutein biosynthesis among different tissues.