

1 **SMRT sequencing generates the chromosome-scale**
2 **reference genome of tropical fruit mango,**
3 ***Mangifera indica***

4
5 **Wei Li^{1*}, Xun-Ge Zhu^{2,3*}, Qun-Jie Zhang^{1*}, Kui Li^{4,5*}, Dan Zhang¹, Cong Shi^{2,3},**
6 **Li-Zhi Gao^{1,2}**

7
8
9 ¹Institution of Genomics and Bioinformatics, South China Agricultural University,
10 Guangzhou 510642, China

11 ²Plant Germplasm and Genomics Center, Germplasm Bank of Wild Species in
12 Southwestern China, Kunming Institute of Botany, Chinese Academy of Sciences,
13 Kunming 650204, China

14 ³University of Chinese Academy of Sciences, Beijing 100039, China

15 ⁴School of Life Sciences, Nanjing University, Nanjing 210093, China

16 ⁵Novogene Bioinformatics Institute, 100083 Beijing, China

17
18
19 *These authors contributed equally to this work.

20
21
22 Correspondence and requests for materials should be addressed to L.-Z.G. (email:
23 Lgaogenomics@163.com)

24

25

26 **Abstract**

27 Mango (*Mangifera indica*), a member of the family Anacardiaceae, is one of the
28 world's most popular tropical fruits. Here we sequenced the variety, "Hong Xiang Ya",
29 and generated a 371.6-Mb mango genome assembly with 34,529 predicted
30 protein-coding genes. Aided with the published genetic map, for the first time, we
31 assembled the *M. indica* genome to the chromosomes, and finally about 98.77% of the
32 genome assembly was anchored to 20 pseudo-chromosomes. The availability of the
33 chromosome-length genome assembly of *M. indica* will provide novel insights into
34 genome evolution, understand the genetic basis of specialized phytochemical
35 composites relevant to fruit quality, and enhance allele mining in genomics-assisted
36 breeding for mango genetic improvement.

37

38

39 **Background & Summary**

40 Mangos, renowned as “king of fruits”, are native to South and Southeast Asia but are
41 now enjoyed all over the world¹. The worldwide production volume of mangos,
42 mangosteens, and guavas reached 50.65 million metric tons in 2017, an increase from
43 46.5 million metric tons in the 2016, which globally grades as the fifth most produced
44 fruit crop. Much of the world’s mangos are produced in the Asia Pacific region, such as
45 India, China and Thailand (<http://www.fao.org/faostat/>). The genus *Mangifera*, which
46 belong to the Anacardiaceae family, comprise about 50 species². While other
47 *Mangifera* species may also yield edible fruits of mangos, almost all cultivated
48 mangoes come from only one species, *Mangifera indica*³. This species possesses a
49 long domestication and cultivation history of over 4,000 years in the Indo-Burmese
50 and Southeast Asia regions, and then spread to other areas of the world since the 14th
51 century^{4,5}. This famous fruit tree is extensively grown in tropical regions and extend
52 to subtropical regions in the world nowadays.

53 While some mango fruits are often processed into various kinds of products, for
54 example, nectar, juice and jam, the majority are predominantly consumed in fresh¹.
55 As many other Anacardiaceae plants, mangos contain phenolic compounds that may
56 stimulate interaction dermatitis, an undesired characteristic for numerous users of
57 fresh mango fruits⁶. However, this fruit with gorgeous appearance and unusual flavors
58 has attracted an increasingly large number of world consumers. Nevertheless, the
59 biosynthesis of these compounds remain largely unresolved up to date. Furthermore,
60 although mango cultivars have heavily relied on vegetative propagation, recent
61 decades have witnessed huge efforts in US, China, Australia and other countries,
62 through conventional cross-breeding programs. The generation of a large number of
63 mango varieties have progressively speeded its wide-reaching distribution in the
64 world⁷. Although some cytogenetics data⁸, genetic mapping^{9,10} and transcriptomics
65 data^{11,12} are publicly available, the lack of chromosome-level high-quality mango
66 reference genome sequences has seriously hampered our understanding of the genetic
67 basis of specialized phytochemical composites and allele mining in genomics-assisted
68 breeding for mango improvement.

69 The completion of chromosome-scale genome assembly of *M. indica* is able to
70 build the foundation for the discovery and application of desired agronomic traits that
71 is of great interest in modern genetic improvement program. Here, we report the first *de*

72 *novo* assembled chromosome-level genome of *M. indica* using SMRT sequencing
73 technology with the assistance of the published genetic map. The obtained genome
74 assembly will greatly help to obtain novel insights into the genome evolution and
75 metabolic biosynthesis of important compounds relevant to fruit quality and enhance
76 the genomics-based trait improvement and mango germplasm utilization.

77 **Methods**

78 **Sample collection, library construction and sequencing.** An individual plant of the
79 *M. indica* variety, “Hong Xiang Ya”, grown in Xishuangbanna City, Yunnan Province,
80 China, was collected for the genome sequencing. Fresh and healthy leaves were
81 harvested and immediately frozen in liquid nitrogen, followed by storage at -80°C in
82 the laboratory prior to DNA extraction. High-quality genomic DNA was isolated using
83 a modified CTAB method¹³ for both Illumina and Pacbio sequencing. The quantity and
84 quality of the DNA sample were examined using a NanoDrop 2000 spectrophotometer
85 (NanoDrop Technologies, Wilmington, DE) and electrophoresis on a 0.8% agarose gel,
86 respectively. For Illumina sequencing, the extracted genomic DNA was fragmented
87 using S220 Focused-ultrasonicator system (Covaris Inc, USA). One paired end library
88 was constructed following the Illumina’s instructions and sequenced on Illumina
89 HiSeq4000 platform. For Pacbio sequencing, a 40-kb SMRTbell DNA library was
90 prepared and sequenced on PacBio Sequel II platform with one SMRT cell. Finally, a
91 total of 87.67 Gb short sequencing data and 151.09 Gb Pacbio data were generated,
92 respectively (**Table 1**).

93
94 **Estimation of genome size and heterozygosity.** The genome size of mango was
95 estimated using *k*-mer frequency distribution generated from short reads. Jellyfish¹⁴
96 was used to calculate 17-mer abundance, and genome size of mango was then
97 estimated using the formula: Genome size = *k*-mer_num/peak_depth, where
98 *k*-mer_number is the total number of *k*-mer and peak_depth is the peak value of *k*-mer
99 frequency distribution. In this study, a total of 43,154,105,487 *k*-mer were counted,
100 while the main peak occurred at depth 111 corresponds to unique haploid sequences.
101 The genome size of mango was estimated to be ~389 Mb. A small peak was also
102 detected at 1/2 peak depth corresponding to considerable heterozygous fractions (**Fig.**
103 **1**). Genomescope¹⁵ was then used to estimate the heterozygosity level of mango. The
104 *k*-mer histogram generated by Jellyfish¹⁴ was subjected to Genomescope¹⁵
105 (<http://qb.cshl.edu/genomescope/>) and heterozygosity estimate for the mango genome
106 was in the range of ~1.84-1.85%.

107

108 ***De novo genome assembly.*** In an attempt to minimize the effect of high
109 heterozygosity, Falcon and Falcon-unzip¹⁶ were used to perform the genome assembly.
110 The longest 55× subreads were selected as seed reads to perform interactive error
111 correction, Falcon was used to obtain primary contigs (p-contigs). The p-contigs were
112 then phased using Falcon-Unzip. Two subsets of contigs were generated, including the
113 primary contigs (p-contigs) and the haplotigs, which represent divergent haplotypes in
114 the assembly. Both p-contigs and haplotigs were polished as follows: firstly, all the
115 pacbio reads were aligned against the assembly using pbalign
116 (<https://github.com/PacificBiosciences/pbalign>). The output files were fed to Arrow
117 (<https://github.com/PacificBiosciences/GenomicConsensus>) implemented in
118 SMRTLink to polish the assembly. Next, the Illumina data from short libraries were
119 aligned to the polished assembly using BWA¹⁷ with default parameters, and then,
120 Pilon¹⁸ was used for sequence assembly refinement based upon these alignments. Two
121 rounds of Pilon was performed to correct the single-base errors and small indels in the
122 assembly. To remove the allelic contigs retained in primary assembly which is
123 deleterious for downstream analysis, Purge Haplotigs¹⁹ was used to identify and
124 reassign allelic contigs. Briefly, the pacbio raw reads were mapped to the p-contigs
125 using Minimap2²⁰. The resulting BAM file was used to generate a read-depth
126 histogram. The collapsed haplotype contigs will fall into the 1×read-depth peak,
127 whereas the allelic contigs will result in half the read-depth. Finally, we generated a
128 primary assembly with a total length of ~372 Mb which spanned 95.37% of the
129 genome size estimated by *k*-mer analysis (**Table 2**). The assembly comprised of 120
130 contigs with an N50 size of 4.82 Mb (**Table 2**). We also generated a combined 168.20
131 Mb of haplotype-resolved sequence, with an N50 of 458.69 Kb and a maximum
132 length of 2,007.74 Kb (**Table 2**). The published genetic map⁹ was also used to anchor
133 scaffolds to chromosomes with ALLMAPS²¹. A total of 112 contigs were anchored to
134 20 pseudochromosomes (**Table 2**). The anchored contigs were 367.05 Mb in size,
135 occupying 98.77% of the genome (**Table 2**). The chromosome lengths of the mango

136 genome varied from ~12 Mbp (Chr10) to ~28 Mbp (Chr19) with an N50 size of ~19
137 Mbp (**Table 3; Fig. 2**).

138

139 **Transposable element annotation.** A mango-specific repeat library was constructed
140 following the instructions of “Repeat Library Construction-Advanced”
141 (http://weatherby.genetics.utah.edu/MAKER/wiki/index.php/Repeat_Library_Construction-Advanced)
142 in MAKER-P pipeline²². Firstly, the genome assembly was searched
143 for miniature inverted transposable elements (MITEs) using MITE-Hunter²³. The
144 identified MITEs were manually checked for TSD and TIR. Long terminal repeat
145 retrotransposons (LTR-RTs) were identified using LTRharvest²⁴ (implemented in
146 GenomeTools²⁵) and LTR_Finder²⁶. LTR_retriever pipeline²⁷ was applied to integrate
147 the results of LTR_Finder²⁶ and LTRharvest²⁴ to efficiently remove false positive
148 elements. Repeatmodeler (<http://www.repeatmasker.org/RepeatModeler/>), which can
149 automatically execute two core *de novo* repeat finding programs, including RECON²⁸
150 and RepeatScout²⁹, was also used to construct repeat library. The elements identified
151 as unknown were searched against the transposase database. An in-house perl script
152 was used to include the sequences matching transposase into the relevant superfamily.
153 All the identified repeat libraries were searched against a plant protein database where
154 proteins from transposons were excluded. Elements with significant hits were filtered
155 using ProtExcluder²². The remained elements were merged to form a comprehensive
156 TE library. Repeatmasker³⁰ was used for the annotation of repetitive sequence. We
157 also identified the tandem repeats using the Tandem Repeat Finder (TRF) package³¹
158 and the non-interspersed repeat sequences using RepeatMasker³⁰ with NCBI search
159 engine. Results showed that approximately 47.94% of the mango genome consists of
160 transposable elements (TEs). LTR retrotransposons were the most abundant TE type,
161 occupying roughly 23.08% of the mango genome (**Table 4**). Most LTRs were
162 LTR/*Gypsy* elements, which occupied 9.16% of the genome (**Table 4**). In addition,
163 18.12% of the mango genome comprised unclassified repetitive sequences (**Table 4**).

164

165 Simple sequence repeats (SSRs) were identified in the mango genome using the
166 MISA³² perl script with the settings: monomer (one nucleotide, $n \geq 12$), dimer (two
167 nucleotides, $n \geq 6$), trimer (three nucleotides, $n \geq 4$), tetramer (four nucleotides, $n \geq 3$),
168 pentamer (five nucleotides, $n \geq 3$), and hexamer (six nucleotides, $n \geq 3$). In total,
169 284,679 SSRs were found, which constitutes 1.1% (4.07 Mb) of the genome (**Table 5**).
170 Among the SSRs, monomer were most dominant (38% of the total SSRs), followed
171 by dimer (26%), tetramer (16%), trimer (11%), pentamer (7%), and hexamer (3%),
172 respectively (**Table 5**).

173

174 **Gene prediction and functional annotation.** Three independent approaches,
175 including the *de novo* method, the homology-based method and the EST-aided
176 method, were used to perform gene prediction. Augustus³³ and SNAP³⁴ were used to
177 perform the *de novo* prediction, with two rounds of iterative training. The protein
178 sequences from *Arabidopsis thaliana* (TAIR10)³⁵, *Anacardium occidentale*
179 (phytozome v12), *Acer yangbiense*³⁶ and *Citrus sinensis*³⁷, were mapped to the
180 genome by Exonerate³⁸, using the Protein2Genome model. For RNA-seq aided gene
181 annotation, the transcriptomic reads of five tissues, including seeds, mesocarps, leaves,
182 flowers, and exocarps, were downloaded from NCBI SRA database under accession
183 number SRR2163402- SRR2163406. Paired-end raw reads were trimmed using
184 Trimmomatic³⁹ to remove adaptors, reads with >3% N and low-quality reads. The
185 quality-filtered reads were subjected to Trinity⁴⁰ to perform *de novo* transcriptome
186 assembly. The resulting transcripts were then aligned to the soft-masked mango
187 genome using GMAP⁴¹ and BLAT⁴². The potential gene structures were iteratively
188 refined using PASA (Program to Assemble Spliced Alignments)⁴³. Finally,
189 EVidenceModeler⁴⁴ was used to integrate the predictions and generate a consensus
190 gene set. Weights of evidences were manually set as: *ab initio* predictions, Augustus
191 = 1, SNAP = 1; protein alignments, Exonerate = 4; EST-aided, PASA = 8. The gene
192 set was refined with TransposonPSI⁴⁴ to remove the TE-related gene model. Gene
193 models with premature termination and/or consisting of fewer than 50 amino acids
194 were also discarded. A total of 34,529 protein-coding genes with an average length of

195 3,295 bp were identified in the mango genome (**Table 6**). The average CDS length
196 and exon number per gene were 1,143 and 5.6, respectively (**Table 6**).

197 The predicted genes were searched against Swiss-Prot⁴⁵ database using BLASTP
198 (e-value cutoff of 10^{-5}). The motifs and domains within gene models were identified
199 using InterProScan⁴⁶. Gene Ontology terms for each gene were directly retrieved from
200 the corresponding InterPro entry. KAAS⁴⁷ web server was applied to perform
201 pathway analysis. All protein sequences were scanned with PfamScan⁴⁸ with Pfam-A
202 database⁴⁹. A total of 90.71% gene models showed significant similarities to
203 sequences in the public databases (**Table 7**).

204

205 **Non-coding RNA gene annotation.** The five different types of non-coding RNA
206 genes, including transfer RNA genes (tRNA), ribosomal RNA genes (rRNA), small
207 nucleolar RNA genes (snoRNAs), small nuclear RNA genes (snRNAs) and
208 microRNA genes (miRNAs), were predicted in the mango genome. The tRNA genes
209 were identified using tRNAscan-SE⁵⁰. RNAmmer⁵¹ was used to predict rRNA genes
210 and their subunits. For the identification of snoRNA, snoScan⁵² was used with the
211 yeast rRNA methylation sites and yeast rRNA sequences provided by the snoScan
212 distribution. Both snRNA genes and miRNA genes were identified by INFERNAL⁵³
213 software against the Rfam database⁵⁴. In total, we identified 598 tRNA genes, 45
214 rRNA genes, 47 snoRNA genes, 200 snRNA genes, and 235 miRNA genes,
215 respectively (**Table 8**).

216

217 **Data Records**

218 All the raw sequencing reads have been deposited to BIG Genome Sequence Archive
219 database under accession number PRJCA002248. The genome assembly and genome
220 annotation are also available at BIG Genome Warehouse under accession number
221 PRJCA002248.

222

223 **Technical Validation**

224 **Assessment of the genome assembly.** First, high-quality reads from NGS sequencing
225 were mapped to the genome assembly using BWA¹⁷. Our results revealed that nearly
226 88.19% Illumina reads were mapped to the genome assembly, among which 84.08%
227 were properly mapped (**Table 9**). Second, the genome assembly was checked with
228 benchmarking universal single-copy orthologs (BUSCO)⁵⁵ from the Embryophyta
229 lineage. Results showed a total of 1343 core orthologs (93.3%) could be found in the
230 mango genome. Third, the RNA sequencing reads were assembled using Trinity⁴⁰ and
231 then aligned back to the assembled genome using GMAP⁴¹. Approximately 84.71% of
232 the transcripts could be mapped to the genome (**Table 9**). Finally, we assessed the
233 mango genome using the LTR Assembly Index (LAI)⁵⁶, which evaluate the quality of
234 assembly by the amount of identifiable intact LTR retroelements. The LAI score of
235 mango is 13.05 (**Fig. 3**), indicating a high quality of genome assembly (draft quality,
236 with LAI score less than 10; reference quality, with LAI score ranges from 10 to 20;
237 and gold quality, with LAI score greater than 20).

238

239 **Improvement of gene annotation quality.** To improve the quality of gene prediction,
240 we performed self-training with Augustus and SNAP. RNA-seq reads were *de novo*
241 assembled using Trinity and refined with PASA to produce additional genome-guided
242 transcriptome assemblies. Manual curation was performed with the training set, genes
243 were retained if: (1) they have the complete gene structure without inner stop codons;
244 (2) they have multiple exons and the CDS length exceed 800 bp. CD-Hit⁵⁷ was used
245 to remove the training set with over 70% sequence similarity. Finally, a total of 2,738
246 high-quality gene models were filtered which served as the training sets. For Augustus,
247 the training sets were randomly divided into two parts, one with 2,238 gene models for
248 training Augustus and the other with 500 gene models for assessing the accuracy of
249 gene model. The protein sequences of mango were downloaded from National Center
250 for Biotechnology Information (NCBI) database and plasmid-related proteins were
251 removed. The protein sequences were then aligned against our gene predictions using
252 BLAST program with an E-value cutoff of 1×10^{-10} . Only hits with coverage $\geq 80\%$
253 and identity $\geq 30\%$ were retained. Results showed that 88.16% of the proteins were

254 supported by our gene predictions, indicating that the annotated mango genes were of
255 high quality (**Table 9**).

256

257 **Code availability**

258 All the bioinformatics tools/packages used in this research described below along with
259 their versions, settings and parameters.

260 **(1) Jellyfish:** version 2.2.10, -m 17 -s 200000M -t 20; **(2) GenomeScope:**
261 k-mer_length=17, read_length=150; **(3) Falcon:** version 0.3.0, genome_size =
262 380000000, seed_coverage = 30, length_cutoff_pr = 5000, max_diff = 100, max_cov
263 = 100; **(4) Falcon-Unzip:** version 0.3.0, default parameters; **(5) Pbalign:** version
264 0.3.1, --nproc 20; **(6) Arrow:** version 2.2.2, -j 50 --maskRadius 3; **(7) BWA:** version
265 0.7.12-r1039, default parameters; **(8) Pilon:** version 1.23, --fix snps,indels; **(9)**
266 **Minimap2:** version 2.17-r974-dirty, -t 20 -ax map-pb --secondary=no; **(10) Purge**
267 **Haplotigs:** -l 40 -m 130 -h 200; **(11) ALLMAPS:** default parameters; **(12)**
268 **MITE-Hunter:** -n 5 -P 1 -S 12345678 -c 20; **(13) GenomeTools:** version 1.5.10, gt
269 ltrharvest -similar 85 -vic 10 -seed 20 -seqids yes -minlenltr 100 -maxlenltr 7000
270 -mintsd 4 -maxtsd 6 -motif TGCA -motifmis 1; **(14) LTR_Finder:** version 1.07, -D
271 15000 -d 1000 -L 7000 -l 100 -p 20 -C -M 0.9; **(15) LTR_retriever:** default
272 parameters; **(16) RECON:** default parameters; **(17) RepeatScout:** version 1.05,
273 default parameters; **(18) RepeatMasker:** version 1.332, -e ncbi; **(19) TRF:** version
274 4.09, default parameters; **(20) RepeatModeler:** version open-1.0.11, -e ncbi; **(21)**
275 **MISA:** 1-12 2-6 3-4 4-3 5-3 6-3; **(22) Augustus:** version 2.7, --gff3=on; **(23) SNAP:**
276 version 2006-07-28, default parameters; **(24) Exonerate:** version 2.2.0, --model
277 protein2genome --minintron 20 --maxintron 30000 --showtargetgff --showvulgar 0
278 --showalignment 0 --softmasktarget TRUE --score 100 --percent 70; **(25)**
279 **Trimmomatic:** version 0.32, LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15
280 MINLEN:36; **(26) Trinity:** version 2.8.4, --seqType fq --full_cleanup
281 --min_contig_length 250; **(27) GMAP:** version 2018-07-04, default parameters; **(28)**
282 **BLAT:** version 36, default parameters; **(29) PASA:** --stringent_alignment_overlap

283 30.0 --MAX_INTRON_LENGTH 20000 --TRANSDECODER; (30)
284 **EvidenceModeler:** --segmentSize 100000 --overlapSize 10000; (31) **TransposonPSI:**
285 default parameters; (32) **InterProScan:** version v5.10–50.0, -goterms; (33)
286 **PfamScan:** default parameters; (34) **tRNAscan-SE:** version 2.0, default parameters;
287 (35) **RNAmmer:** version 1.2, -S euk -m lsu,ssu,tsu; (36) **snoscan:** version 0.9.1,
288 default parameters; (37) **INFERNAL:** version 1.1.2, default parameters; (38)
289 **BUSCO:** version 3.0.2, default parameters; (39) **LAI:** default parameters; (40)
290 **CD-Hit:** version 4.8.1, -c 0.7.

291

292 **References**

- 293 1 Tharanathan, R., Yashoda, H. & Prabha, T. Mango (*Mangifera indica* L.), “The
294 king of fruits”—An overview. *Food Reviews International* **22**, 95-123 (2006).
- 295 2 Wannan, B. Analysis of generic relationships in Anacardiaceae.
296 *Blumea-Biodiversity, Evolution and Biogeography of Plants* **51**, 165-195
297 (2006).
- 298 3 Kostermans, A. J. G. *The mangoes: Their botany, nomenclature, horticulture*
299 *and utilization*. (Academic Press, 2012).
- 300 4 Mehrotra, R., Dilcher, D. & Awasthi, N. A Palaeocene *Mangifera*-like leaf
301 fossil from India. *Phytomorphology* **48**, 91-100 (1998).
- 302 5 Sawangchote, P., Grote, P. J. & Dilcher, D. L. Tertiary leaf fossils of *Mangifera*
303 (*Anacardiaceae*) from Li Basin, Thailand as examples of the utility of leaf
304 marginal venation characters. *American journal of botany* **96**, 2048-2061
305 (2009).
- 306 6 Schulze-Kaysers, N., Feuereisen, M. & Schieber, A. Phenolic compounds in
307 edible species of the *Anacardiaceae* family—a review. *RSC Advances* **5**,
308 73301-73314 (2015).
- 309 7 Knight, R. J. & Schnell, R. J. Mango introduction in Florida and
310 the ‘haden’ cultivar’s significance to the modern industry. *Economic botany* **48**,
311 139-145 (1994).
- 312 8 Mukherjee, S. K. Mango: its allopolyploid nature. *Nature* **166**, 196-197 (1950).
- 313 9 Luo, C. *et al.* Construction of a high-density genetic map based on large-scale
314 marker development in mango using specific-locus amplified fragment
315 sequencing (SLAF-seq). *Frontiers in plant science* **7**, 1310 (2016).
- 316 10 Kuhn, D. N. *et al.* Genetic map of mango: a tool for mango breeding. *Frontiers*
317 *in plant science* **8**, 577 (2017).
- 318 11 Tafolla-Arellano, J. C. *et al.* Transcriptome analysis of mango (*Mangifera*
319 *indica* L.) fruit epidermal peel to identify putative cuticle-associated genes.
320 *Scientific reports* **7**, 46163 (2017).

- 321 12 Sivankalyani, V. *et al.* Transcriptome dynamics in mango fruit peel reveals
322 mechanisms of chilling stress. *Frontiers in plant science* **7**, 1579 (2016).
- 323 13 Porebski, S., Bailey, L. G. & Baum, B. R. Modification of a CTAB DNA
324 extraction protocol for plants containing high polysaccharide and polyphenol
325 components. *Plant Molecular Biology Reporter* **15**, 8-15 (1997).
- 326 14 Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel
327 counting of occurrences of k-mers. *Bioinformatics* **27**, 764-770,
328 doi:10.1093/bioinformatics/btr011 (2011).
- 329 15 Vurture, G. W. *et al.* GenomeScope: Fast reference-free genome profiling from
330 short reads. *Bioinformatics* **33** (2017).
- 331 16 Chin, C. S., Peluso, P. & Sedlazeck, F. J. Phased diploid genome assembly with
332 single-molecule real-time sequencing. **13**, 1050-1054,
333 doi:10.1038/nmeth.4035 (2016).
- 334 17 Li, H. Li, H.: Aligning sequence reads, clone sequences and assembly contigs
335 with BWA-MEM. arXiv (1303.3997). **1303** (2013).
- 336 18 Walker, B. J. *et al.* Pilon: an integrated tool for comprehensive microbial variant
337 detection and genome assembly improvement. *PLoS One* **9**, e112963,
338 doi:10.1371/journal.pone.0112963 (2014).
- 339 19 Roach, M. J., Schmidt, S. A. & Borneman, A. R. Purge Haplotigs: allelic contig
340 reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics* **19**,
341 460, doi:10.1186/s12859-018-2485-7 (2018).
- 342 20 Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*
343 **34**, 3094-3100, doi:10.1093/bioinformatics/bty191 (2018).
- 344 21 Tang, H. *et al.* ALLMAPS: robust scaffold ordering based on multiple maps.
345 *Genome Biology* **16**, 3.
- 346 22 Campbell, M. S., Law, M., Holt, C., Stein, J. C. & Yandell, M. MAKER-P: A
347 Tool Kit for the Rapid Creation, Management, and Quality Control of Plant
348 Genome Annotations. *Plant Physiology* **164**, 513 (2013).

- 349 23 Han, Y. & Wessler, S. R. MITE-Hunter: a program for discovering miniature
350 inverted-repeat transposable elements from genomic sequences. *Nucleic Acids*
351 *Res* **38**, e199, doi:10.1093/nar/gkq862 (2010).
- 352 24 Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible
353 software for de novo detection of LTR retrotransposons. *BMC Bioinformatics* **9**,
354 18, doi:10.1186/1471-2105-9-18 (2008).
- 355 25 Gremme, G., Steinbiss, S. & Kurtz, S. GenomeTools: a comprehensive
356 software library for efficient processing of structured genome annotations.
357 *IEEE/ACM Trans Comput Biol Bioinform* **10**, 645-656,
358 doi:10.1109/tcbb.2013.68 (2013).
- 359 26 Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of
360 full-length LTR retrotransposons. *Nucleic Acids Res* **35**, W265-268,
361 doi:10.1093/nar/gkm286 (2007).
- 362 27 Ou, S. & Jiang, N. LTR_retriever: A Highly Accurate and Sensitive Program
363 for Identification of Long Terminal Repeat Retrotransposons. *Plant Physiol*
364 **176**, 1410-1422, doi:10.1104/pp.17.01310 (2018).
- 365 28 Bao, Z. & Eddy, S. R. Automated *de novo* identification of repeat sequence
366 families in sequenced genomes. *Genome Res* **12**, 1269-1276,
367 doi:10.1101/gr.88502 (2002).
- 368 29 Price, A. L., Jones, N. C. & Pevzner, P. A. *De novo* identification of repeat
369 families in large genomes. *Bioinformatics* **21 Suppl 1**, i351-358,
370 doi:10.1093/bioinformatics/bti1018 (2005).
- 371 30 Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive
372 elements in genomic sequences. *Curr Protoc Bioinformatics* **Chapter 4**, Unit
373 4.10, doi:10.1002/0471250953.bi0410s25 (2009).
- 374 31 Benson, G. Tandem repeats finder: a program to analyze DNA sequences.
375 *Nucleic Acids Res* **27**, 573-580, doi:10.1093/nar/27.2.573 (1999).
- 376 32 Thiel, T., Michalek, W., Varshney, R. K. & Graner, A. Exploiting EST
377 databases for the development and characterization of gene-derived

- 378 SSR-markers in barley (*Hordeum vulgare* L.). *Theor Appl Genet* **106**, 411-422,
379 doi:10.1007/s00122-002-1031-0 (2003).
- 380 33 Stanke, M., Steinkamp, R., Waack, S. & Morgenstern, B. AUGUSTUS: a web
381 server for gene finding in eukaryotes. *Nucleic Acids Res* **32**, W309-312,
382 doi:10.1093/nar/gkh379 (2004).
- 383 34 Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59,
384 doi:10.1186/1471-2105-5-59 (2004).
- 385 35 TheArabidopsisGenomeInitiative. Analysis of the genome sequence of the
386 flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796-815,
387 doi:10.1038/35048692 (2000).
- 388 36 Yang, J. *et al.* De novo genome assembly of the endangered *Acer yangbiense*, a
389 plant species with extremely small populations endemic to Yunnan Province,
390 China. *Gigascience* **8**, doi:10.1093/gigascience/giz085 (2019).
- 391 37 Xu, Q. *et al.* The draft genome of sweet orange (*Citrus sinensis*). *Nature*
392 *Genetics* **45**, 59-66, doi:10.1038/ng.2472 (2013).
- 393 38 Slater, G. S. & Birney, E. Automated generation of heuristics for biological
394 sequence comparison. *BMC Bioinformatics* **6**, 31, doi:10.1186/1471-2105-6-31
395 (2005).
- 396 39 Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for
397 Illumina sequence data. *Bioinformatics* **30**, 2114-2120,
398 doi:10.1093/bioinformatics/btu170 (2014).
- 399 40 Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data
400 without a reference genome. *Nat Biotechnol* **29**, 644-652, doi:10.1038/nbt.1883
401 (2011).
- 402 41 Wu, T. D. & Watanabe, C. K. GMAP: a genomic mapping and alignment
403 program for mRNA and EST sequences. *Bioinformatics* **21**, 1859-1875,
404 doi:10.1093/bioinformatics/bti310 (2005).
- 405 42 Kent, W. J. BLAT--the BLAST-like alignment tool. *Genome Res* **12**, 656-664,
406 doi:10.1101/gr.229202 (2002).

- 407 43 Haas, B. J. *et al.* Improving the Arabidopsis genome annotation using maximal
408 transcript alignment assemblies. *Nucleic Acids Res* **31**, 5654-5666 (2003).
- 409 44 Haas, B. J. *et al.* Automated eukaryotic gene structure annotation using
410 EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome*
411 *Biol* **9**, R7, doi:10.1186/gb-2008-9-1-r7 (2008).
- 412 45 Boeckmann, B. *et al.* The SWISS-PROT protein knowledgebase and its
413 supplement TrEMBL in 2003. *Nucleic Acids Res* **31**, 365-370 (2003).
- 414 46 Jones, P. *et al.* InterProScan 5: genome-scale protein function classification.
415 *Bioinformatics* **30**, 1236-1240, doi:10.1093/bioinformatics/btu031 (2014).
- 416 47 Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. C. & Kanehisa, M. KAAS: an
417 automatic genome annotation and pathway reconstruction server. *Nucleic Acids*
418 *Res* **35**, W182-185, doi:10.1093/nar/gkm321 (2007).
- 419 48 Li, W. *et al.* The EMBL-EBI bioinformatics web and programmatic tools
420 framework. *Nucleic Acids Res* **43**, W580-584, doi:10.1093/nar/gkv279 (2015).
- 421 49 Punta, M. *et al.* The Pfam protein families database. *Nucleic Acids Res* **40**,
422 D290-301, doi:10.1093/nar/gkr1065 (2012).
- 423 50 Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of
424 transfer RNA genes in genomic sequence. *Nucleic Acids Res* **25**, 955-964
425 (1997).
- 426 51 Lagesen, K. *et al.* RNAmmer: consistent and rapid annotation of ribosomal
427 RNA genes. *Nucleic Acids Res* **35**, 3100-3108, doi:10.1093/nar/gkm160 (2007).
- 428 52 Schattner, P., Brooks, A. N. & Lowe, T. M. The tRNAscan-SE, snoscan and
429 snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids*
430 *Res* **33**, W686-689, doi:10.1093/nar/gki366 (2005).
- 431 53 Nawrocki, E. P., Kolbe, D. L. & Eddy, S. R. Infernal 1.0: inference of RNA
432 alignments. *Bioinformatics* **25**, 1335 (2009).
- 433 54 Griffiths-Jones, S. *et al.* Rfam: annotating non-coding RNAs in complete
434 genomes. *Nucleic Acids Res* **33**, D121-124, doi:10.1093/nar/gki081 (2005).
- 435 55 Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov,
436 E. M. BUSCO: assessing genome assembly and annotation completeness with

437 single-copy orthologs. *Bioinformatics* **31**, 3210-3212,
438 doi:10.1093/bioinformatics/btv351 (2015).

439 56 Ou, S., Chen, J. & Jiang, N. Assessing genome assembly quality using the LTR
440 Assembly Index (LAI). *Nucleic Acids Res* **46**, e126, doi:10.1093/nar/gky730
441 (2018).

442 57 Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large
443 sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658-1659,
444 doi:10.1093/bioinformatics/btl158 (2006).

445

446

447 **Acknowledgements**

448 This work was supported by Yunnan Innovation Team Project and Natural Science
449 Foundation of Yunnan (to L.-Z.G.).

450

451 **Author contributions**

452 LZG conceived and designed the study; XGZ, CS and DZ contributed to the
453 collection and preparation of the samples; WL and KL performed the genome
454 assembly; WL performed genome annotation; QJZ performed data visualization; WL
455 and LZG drafted the manuscript; LZG revised the manuscript.

456

457 **Additional Information**

458 **Competing interests:** The authors declare no competing interests.

459

460

461 **Figure Legends**

462 **Figure 1. The 17-mer distribution of sequencing reads of the mango genome.** The
463 occurrence of 17-mer was calculated using jellyfish based on the sequencing data from
464 short insert size libraries (insert size \leq 500 bp) of mango. The sharp peak on left with
465 low depths represents the essentially random sequencing errors. The middle and right
466 peaks indicate the heterozygous and homozygous peaks, the depths of which are 55 and
467 111, respectively.

468

469 **Figure 2. The mango genome features.** (A) Circular representation of the 20
470 pseudochromosomes; (B) the distribution of TEs. Purple for DNA TEs, red for
471 Ty3-gypsy LTR-RTs, orange for Ty1-copia LTR-RTs, yellow for unclassified
472 LTR-RTs. and light yellow for other repeats; (C) GC content; (D) number of SSRs per
473 500kb; (E) the distribution of protein-coding genes; (F) the distribution of non-coding
474 RNAs, yellow for tRNAs, orange for miRNAs, red for snRNAs, purple for snoRNAs,
475 and blue for rRNAs.

476

477 **Figure 3. LAI scores in genomic regions of mango.** Each dot represents LAI score
478 with 300-Kb sliding window. The red line indicates the Draft-Reference boundary and
479 the blue line shows the Reference-Gold boundary (Draft, such as Apple v1.0 and
480 Cacao v1.0; Reference, such as Arabidopsis TAIR10; Gold, such as Rice MSUv7).

481

Sequencing technique	Insert Size (bp)	Average Read Length (bp)	Raw Data (Mb)	Raw Sequence Coverage (×)
Next-generation	260	150	87,671	225.38
Pacbio	40,000	20,467	151,092	388.41

482

483 **Table 1. Libraries and read statistics used for the mango genome assembly.** The
484 estimated genome size is ~389 Mb.

485

Features	P-contigs	H-contigs	Pseudochromosomes
Genome size (bp)	371,611,606	168,197,039	367,054,942
Contig number	120	1,190	112
Max contig length (bp)	11,465,205	2,007,740	11,465,205
Contig N50 (bp)	4,823,431	458,693	4,951,396
Contig N90 (bp)	1,742,740	56,071	1,768,123
Scaffold N50 (bp)	4,823,431	458,693	18,777,690
Scaffold N90 (bp)	1,742,740	56,071	13,631,091
Gap number	0	0	92
Gap length (bp)	0	0	9,200
GC content (%)	32.61	32.87	32.62

486

487 **Table 2. Assembly statistics of the mango genome.**

488

Chromosome ID	Contig number	Chromosome length (bp)
1	6	17,272,295
2	8	19,959,718
3	6	18,777,690
4	9	22,014,470
5	7	18,043,982
6	5	13,934,890
7	6	16,887,665
8	6	20,966,627
9	8	22,271,002
10	2	12,445,408
11	6	23,906,801
12	5	15,001,369
13	5	18,750,349
14	3	13,865,489
15	6	13,631,091
16	3	19,594,037
17	2	11,943,369
18	4	21,874,139
19	7	27,760,119
20	8	18,154,432
Unanchored	8	4,565,864
Total	120	371,620,806

489

490 **Table 3. Chromosome lengths of the assembled mango genome.**

491

492

Transposable Elements	Length (bp)	Percentage (%)
DNA transposons elements	16,187,808	4.36
CMC	157,712	0.04
Harbinger	89,828	0.02
hAT	3,270,678	0.88
MuLE	10,873,608	2.93
Helitron	1,713,184	0.46
Others	82,798	0.02
RNA transposons elements	88,061,304	23.70
Non-LTR Retrotransposons	2,297,579	0.62
LINEs	2,297,579	0.62
LTR Retrotransposons	85,763,725	23.08
<i>Copia</i>	24,945,038	6.71
<i>Gypsy</i>	34,039,343	9.16
Caulimovirus	418,105	0.11
Others	26,361,239	7.09
Unclassified elements	67,335,389	18.12
Other Repeats	6,580,619	1.77
Low complexity	906,641	0.24
Simple repeats	5,673,978	1.53
Total	178,165,120	47.94

493

494 **Table 4. Statistics of repeat sequences in the mango genome.**

495

496

Repeat type	Total counts	Total length (bp)	Average length (bp)	Proportion (%)	Frequency (loci/Mb)	Density (bp/Mb)
Monomer	106,814	1,324,805	12	37.52	287.13	3,561.30
Dimer	72,740	1,220,894	17	25.55	195.54	3,281.97
Trimer	31,735	447,825	14	11.15	85.31	1,203.83
Tetramer	45,874	608,212	13	16.11	123.32	1,634.98
Pentamer	20,350	326,085	16	7.15	54.70	876.57
Hexamer	7,166	138,234	19	2.52	19.26	371.60
Total	284,679	4,066,055	14	100	765.27	10,930.26

497

498 **Table 5. Occurrence of simple sequence repeats (SSRs) in the mango genome.**

499

Gene set		Number	Average length (bp)	Average CDS length (bp)	Average intron length (bp)	Average exon per gene
De novo	Augustus	33,485	2,709	1,203	330	5.6
	SNAP	48,490	4,316	801	792	5.4
Homolog	<i>Arabidopsis thaliana</i>	15,268	3,138	1,099	428	5.8
	<i>Anacardium occidentale</i>	117,919	3,860	1,247	479	6.5
	<i>Acer yangbiense</i>	13,434	3,321	1,264	477	5.3
	<i>Citrus sinensis</i>	26,222	3,791	1,320	459	6.4
RNA-seq	PASA	100,052	1,988	1,079	451	3.0
Final set	EVM	34,529	3,295	1,143	465	5.6

500

501 **Table 6. Prediction of protein-coding genes in the mango genome.**

502

503

Database	Number	Percentage (%)
Total	34,529	100
InterPro	31,143	90.19
GO	20,298	58.79
KEGG	8,731	25.29
Pfam	24,846	71.96
Swissprot	25,834	74.82
Unannotated	3,208	9.29

504

505 **Table 7. Functional annotation of the mango genome.**

506

Non-coding RNAs	Number	Average Length (bp)	Total Length (bp)
tRNA	598	74.13	44,332
rRNA (8S)	41	113.29	4,645
rRNA (18S)	1	1,808	1,808
rRNA (28S)	3	5,321	15,963
snoRNA	47	94.87	4,459
snRNA	200	127.04	25,408
miRNA	235	121.57	28,570

507

508 **Table 8. Statistics for non-coding RNA genes in the mango genome.**

509

	Methods	Total	Aligned	Percentage (%)
Genome Assembly	Reads mapping			
	PE reads	584,474,718	515,432,238	88.19
	Transcripts of mango			
	EST	174,130	147,503	84.71
	BUSCOs from Embryophyta lineage			
	Complete	1,440	1,343	93.26
	Duplicated	1,440	237	16.46
	Fragmented	1,440	29	2.01
	Missing	1,440	68	4.72
Genome Annotation	Proteins from NCBI	532	469	88.16

510

511 **Table 9. Quality assessment of the mango genome assembly and genome**
 512 **annotation.**





