

GigaScience

The genome draft of the coconut (*Cocos nucifera*)

--Manuscript Draft--

Manuscript Number:	GIGA-D-17-00038R2	
Full Title:	The genome draft of the coconut (<i>Cocos nucifera</i>)	
Article Type:	Data Note	
Funding Information:	International Science and Technology Cooperation projects of Hainan Province (No. KJHZ2014-24)	Dr Yaodong Yang
	Hainan Natural Science Foundation (313058)	Dr Wei Xia
	the fundamental Scientific Research Funds for Chinese Academy of Tropical Agriculture Sciences (1630032012044)	Dr Yaodong Yang
	the fundamental Scientific Research Funds for Chinese Academy of Tropical Agriculture Sciences (1630052014002)	Dr Yaodong Yang
	the fundamental Scientific Research Funds for Chinese Academy of Tropical Agriculture Sciences (1630052015050)	Dr Yong Xiao
	The major Technology Project of Hainan (ZDZX2013023-1)	Dr Ming Peng
	the fundamental Scientific Research Funds for Chinese Academy of Tropical Agriculture Sciences (1630152017019)	Dr Yaodong Yang
	the fundamental Scientific Research Funds for Chinese Academy of Tropical Agriculture Sciences (1630152016006)	Dr Yong Xiao
	Central Public-interest Scientific Institution Basal Research Fund for Innovative Research Team Program of CATAS (17CXTD-28)	Dr Yaodong Yang
Abstract:	<p>Background Coconut palm (<i>Cocos nucifera</i>, $2n = 32$), a member of genus <i>Cocos</i> and family <i>Arecaceae</i> (<i>Palmaceae</i>), is an important tropical fruit and oil crop. Currently, coconut palm is cultivated in 93 countries, including Central and South America, East and West Africa, Southeast Asia and the Pacific island, with a total growth area of more than 12 million hectares (www.fao.org/faostat/en/). Coconut palm is generally classified into two main categories: "Tall" (flowering 8-10 years after planting) and "Dwarf" (flowering 4-6 years after planting), based on morphological characteristics and breeding habits. This <i>Palmae</i> species has a long growth period before reproductive years which hinders conventional breeding progress. In spite of initial successes, improvements made by conventional breeding have been very slow. In the present study, we obtained <i>de novo</i> sequences of <i>Cocos nucifera</i> genome: a major genomic resource which could be used to facilitate molecular breeding in <i>Cocos nucifera</i> and accelerating the breeding process in this important crop.</p> <p>Findings A total of 419.67 gigabases (Gb) of raw reads were generated by the IlluminaHiSeq 2000 platform using a series of paired-end and mate-pair libraries, covering the predicted <i>Cocos nucifera</i> genome length (2.42Gb, variety "Hainan Tall") to an estimated $173.32\times$ read depth. A total scaffold length of 2.20 Gb was generated ($N50 = 418$ Kb), representing 90.91% of the genome. The coconut genome was predicted to harbor 28,039 protein-coding genes, which is less than in <i>Phoenix dactylifera</i> (PDK30 variety: 28,889), <i>Phoenix dactylifera</i> (DPV01 variety: 41,660) and <i>Elaeis guineensis</i> (34,802). BUSCO evaluation demonstrated the obtained scaffold sequences covered 90.8% of the coconut genome, and that the genome annotation was 74.1% complete. Genome annotation results revealed that 72.75% of the coconut</p>	

	<p>genome was consisted of transposable elements. of which long-terminal repeat retrotransposons elements (LTRs) accounted for the largest proportion (92.23%). Comparative analysis of the antiporter gene family and ion channel gene families between <i>C. nucifera</i> and <i>Arabidopsis thaliana</i> indicated that significant gene expansion may occurred in coconut involving Na⁺/H⁺ antiporter, Carnitine/acylcarnitine translocase, Potassium-dependent sodium-calcium exchanger, and potassium channel genes.</p> <p>Conclusions</p> <p>Despite its agronomic importance, <i>C. nucifera</i> is still under-studied. In this report, we made an attempt to construct a draft genome of <i>C. nucifera</i> and provide an enormous amount of genomic information that will facilitate future functional genomics and molecular assisted breeding in this crop species.</p>
Corresponding Author:	Yaodong Yang, Ph.D Coconut Research Institute Wenchang City, Hainan CHINA
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Coconut Research Institute
Corresponding Author's Secondary Institution:	
First Author:	Yong Xiao
First Author Secondary Information:	
Order of Authors:	<p>Yong Xiao</p> <p>Pengwei Xu</p> <p>Haikuo Fan</p> <p>Luc Baudouin</p> <p>Wei Xia</p> <p>Stéphanie Bocs</p> <p>Junyang Xu</p> <p>Qiong Li</p> <p>Anping Guo</p> <p>Lixia Zhou</p> <p>Jing Li</p> <p>Yi Wu</p> <p>Zilong Ma</p> <p>Alix Armero</p> <p>Auguste Emmanuel Issali</p> <p>Na Liu</p> <p>Ming Peng</p> <p>Yaodong Yang</p>
Order of Authors Secondary Information:	
Response to Reviewers:	<p>Response to editor and reviewers</p> <p>Dear editor</p> <p>Thank you very much for your crucial comments for our manuscript entitled "The genome draft of the coconut (<i>Cocos nucifera</i>)" (GIGA-D-17-00038). We have made some revisions to the ms based on your comments. We removed revision mode and</p>

	<p>upload the clean manuscript without any tracking of changes. Meanwhile, we cited the reference database “Xiao Y, Xu P, Fan H, Baudouin L, Xia W, Bocs S et al. Supporting data for The genome draft of coconut (Cocos nucifera). Gigascience Database. 2017” in the revised manuscript. We hope the revised version can meet the requirement of “GigaScience”.</p> <p>Sincerely yours,</p> <p>Yaodong Yang</p> <p>PHONE NUMBER: 0086-898-63330602 FAX NUMBER: 0086-898-63330673 EMAIL: yyang@catas.cn POSTAL ADDRESS: Coconuts Research Institute, Chinese Academy of Tropical Agricultural Sciences, 496 Wenqing Av., Wenchang, Hainan 571339, P.R.China</p>
Additional Information:	
Question	Response
Are you submitting this manuscript to a special series or article collection?	No
<p>Experimental design and statistics</p> <p>Full details of the experimental design and statistical methods used should be given in the Methods section, as detailed in our Minimum Standards Reporting Checklist. Information essential to interpreting the data presented should be made available in the figure legends.</p> <p>Have you included all the information requested in your manuscript?</p>	Yes
<p>Resources</p> <p>A description of all resources used, including antibodies, cell lines, animals and software tools, with enough information to allow them to be uniquely identified, should be included in the Methods section. Authors are strongly encouraged to cite Research Resource Identifiers (RRIDs) for antibodies, model organisms and tools, where possible.</p> <p>Have you included the information requested as detailed in our Minimum Standards Reporting Checklist?</p>	Yes
<p>Availability of data and materials</p> <p>All datasets and code on which the conclusions of the paper rely must be either included in your submission or</p>	Yes

deposited in [publicly available repositories](#) (where available and ethically appropriate), referencing such data using a unique identifier in the references and in the “Availability of Data and Materials” section of your manuscript.

Have you have met the above requirement as detailed in our [Minimum Standards Reporting Checklist](#)?

1 **The genome draft of coconut (*Cocos nucifera*)**

2 Yong Xiao^{1*}, Pengwei Xu^{3*}, Haikuo Fan^{1*}, Luc Baudouin^{4,5*}, Wei Xia¹, Stéphanie Bocs^{4,5}, Junyang

3 Xu³, Qiong Li², Anping Guo², Lixia Zhou¹, Jing Li¹, Yi Wu¹, Zilong Ma², Alix Armero^{4,6}, Auguste

4 Emmanuel Issali⁷, Na Liu³, Ming Peng^{2&}, Yaodong Yang^{1&}

5 ¹Hainan Key Laboratory of Tropical Oil Crops Biology/Coconut Research Institute, Chinese

6 Academy of Tropical Agricultural Sciences, Wenchang, Hainan 571339, P.R.China

7 ²Institute of Tropical Bioscience and Biotechnology, Chinese Academy of Tropical Agricultural

8 Science, Haikou, Hainan 571101, P. R. China

9 ³BGI-Shenzhen, Shenzhen 518083, China

10 ⁴AGAP, Université de Montpellier, CIRAD, INRA, Montpellier Supagro, F-34398 Montpellier,

11 France

12 ⁵CIRAD , UMR AGAP, F-34398, Montpellier France

13 ⁶Montpellier Supagro , UMR AGAP, F-34398, Montpellier France

14 ⁷Station Cocotier Marc Delorme, Centre National De RechercheAgronomique (CNRA)07 B.P. 13,

15 Port Bouet,Côte d'Ivoire

16

17 *The authors have equal contribution to the manuscript

18 &Corresponding author

19 **Yong Xiao:** xiaoyong1980@catas.cn

20 **Pengwei Xu:** xupengwei@genomics.cn

21 **Haikuo Fan:** vanheco@163.com

22 **Luc Baudouin:** luc.baudouin@cirad.fr

23 **Wei Xia:** s aizixiawei@hainu.edu.cn

24 **Stéphanie Bocs:** stephanie.sidibe-bocs@cirad.fr

25 **Junyang Xu:** xujy@genomics.cn

26 **Qiong Li:** liqiong4416@126.com

27 **Anping Guo:** gap211@126.com

28 **Lixia Zhou:** glzz_2009@163.com

29 **Jing Li:** lijing002x@catas.cn

30 **Yi Wu:** wuyi-scuta@163.com

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

31 **Zilong Ma:** mzl900@163.com
32 **Alix Armero:** alix.armero_villanueva@cirad.fr
33 **Auguste Emmanuel Issali:** issaliemma@yahoo.com
34 **Na Liu:** naliu@genomics.cn
35 **Ming Peng:** mmpeng_2000@yahoo.com
36 **Yaodong Yang:** yvjang@catas.cn

38 **Background**

39 Coconut palm (*Cocos nucifera*, $2n = 32$), a member of genus *Cocos* and family Arecaceae
40 (Palmaceae), is an important tropical fruit and oil crop. Currently, coconut palm is cultivated in 93
41 countries, including Central and South America, East and West Africa, Southeast Asia and the Pacific
42 island, with a total growth area of more than 12 million hectares (www.fao.org/faostat/en/). Coconut
43 palm is generally classified into two main categories: “Tall” (flowering 8-10 years after planting) and
44 “Dwarf” (flowering 4-6 years after planting), based on morphological characteristics and breeding
45 habits. This Palmae species has a long growth period before reproductive years which hinders
46 conventional breeding progress. In spite of initial successes, improvements made by conventional
47 breeding have been very slow. In the present study, we obtained *de novo* sequences of *Cocos nucifera*
48 genome: a major genomic resource which could be used to facilitate molecular breeding in *Cocos*
49 *nucifera* and accelerating the breeding process in this important crop.

50 **Findings**

51 A total of 419.67 gigabases (Gb) of raw reads were generated by the Illumina HiSeq 2000 platform
52 using a series of paired-end and mate-pair libraries, covering the predicted *Cocos nucifera* genome
53 length (2.42Gb, variety “Hainan Tall”) to an estimated $173.32\times$ read depth. A total scaffold length of
54 2.20 Gb was generated ($N50 = 418$ Kb), representing 90.91% of the genome. The coconut genome
55 was predicted to harbor 28,039 protein-coding genes, which is less than in *Phoenix dactylifera*
56 (PDK30 variety: 28,889), *Phoenix dactylifera* (DPV01 variety: 41,660) and *Elaeis guineensis*
57 (34,802). BUSCO evaluation demonstrated the obtained scaffold sequences covered 90.8% of the
58 coconut genome, and that the genome annotation was 74.1% complete. Genome annotation results
59 revealed that 72.75% of the coconut genome was consisted of transposable elements. Of which
60 long-terminal repeat retrotransposons elements (LTRs) accounted for the largest proportion (92.23%).

1 61 Comparative analysis of the antiporter gene family and ion channel gene families between *C. nucifera*
2 62 and *Arabidopsis thaliana* indicated that significant gene expansion may occurred in coconut involving
3
4 63 Na⁺/H⁺ antiporter, Carnitine/acylcarnitine translocase, Potassium-dependent sodium-calcium
5
6 64 exchanger, and potassium channel genes.
7

8 65 **Conclusions**

9
10 66 Despite its agronomic importance, *C. nucifera* is still under-studied. In this report, we present a draft
11
12 67 genome of *C. nucifera* and provide genomic information that will facilitate future functional
13
14 68 genomics and molecular assisted breeding in this crop species.
15
16
17 69

18
19 70 **Keywords: Coconut palm, genome, Assembly, Annotation**
20

21 71

22 72 **Data description**

23 73 **Background**

24
25 74 Coconut palm (*Cocos nucifera*, 2n = 32), the only species in genus *Cocos* in the family *Arecaceae*, is
26
27 75 a tropical oil crop and widely cultivated in tropical regions due to its extensive application in
28
29 76 agriculture and industry. Coconut palm is thought to be originated from the Southwest and Western
30
31 77 Pacific region (including the Malay Peninsula and Archipelago, New Guinea, and the Bismarck
32
33 78 Archipelago). At present, this tropical tree crop is distributed across 93 tropical countries [1],
34
35 79 including Central and South American, East and West African, Southeast Asia and the Pacific Islands,
36
37 80 and is grown over 12 million hectares of land (www.fao.org/faostat/en/).
38
39
40

41 81 In China, coconut palm grows in the subtropical regions - Hainan and Yunnan provinces - as an
42
43 82 economic and ornamental plant. Coconut palm is cultivated over approximately 43,000 hectares in
44
45 83 Hainan, with the “Hainan Tall” (HAT) variety covering 36,000 hectares [2]. The HAT coconut needs
46
47 84 eight to ten years to enter its reproductive stage and has a height of 20-30 meters with a medium to
48
49 85 large sized nut. The HAT cultivar is highly tolerant to salt and drought stress, but sensitive to
50
51 86 temperatures below 10 °C. Coconut palm can disseminate through ocean currents: floating nuts sprout
52
53 87 and grow naturally upon washing up on beaches. The ability to adapt to a high salt environment is
54
55 88 closely related to this dissemination feature and to these natural growth conditions. The
56
57 89 morphological characteristics of the HAT cultivar are shown in Figure 1. Here, we present the genome
58
59 90 sequence of the Hainan Tall coconut and an analysis of the antiporter and ion channel gene families,
60
61
62
63
64
65

1 91 relevant to salinity tolerance. As draft genome sequences of coconut relatives (e.g. *Elaeis guineensis*
2 92 [3] and *Phoenix dactylifera* [4, 5]) have previously been reported, we also performed a comparative
3
4 93 analysis between coconut and these relative species for genome assembly and annotation
5
6 94 characteristics.

7
8 95

96 **Data description**

97 **Sample collection and sequencing strategy**

98 The genomic DNA was extracted from the spear leaf of the variety “Hainan Tall” coconut (*Cocos*
99 *nucifera* L. Taxonomy ID: 13894; 19°33’3”N, 110°47’25” E) individual from the coconut garden of
100 the Coconut Research Institute (Wenchang, Hainan province, China) by using the CTAB extraction
101 method [6]. Subsequently, four paired-end (PE) libraries with insert sizes of 170 bp, 500 bp, 450 bp
102 and 800 bp and five mate-pair (MP) libraries with insert sizes of 2 Kb, 5 Kb, 10 Kb, 20 Kb and 40 Kb
103 were constructed using the standard procedure provided by Illumina (San Diego, USA). After library
104 preparation and quality control of the DNA samples, template DNA fragments were hybridized to the
105 surface of the flow cells on an Illumina HiSeq2000 sequencer and amplified to form clusters and then
106 sequenced by following the standard Illumina manual. Finally, we generated 714.67 Gb of raw reads
107 from all constructed libraries. The raw outputs for each sequenced library are summarized in Table 1.
108 Before assembly, the raw reads were pretreated using the following stringent filtering processes via
109 the SOAPfilter (v2.2) [7] software: (1) removed reads with 25% low-quality bases (quality scores \leq
110 7); (2) removed reads with N bases more than 1%; (3) discarded reads with adapter contamination
111 and/or PCR duplicates; (4) removed reads with undersized insert sizes. Finally, 419.08 Gb (estimated
112 173.17 \times read depth) of high-quality sequences were obtained for genome assembly.

113 ***De novo* assembly of short reads of *Cocos nucifera***

114 We used 209.38 Gb clean reads of the short-insert libraries (excluding the 450bp library) to estimate
115 the coconut genome size by k-mer frequency distribution analysis [7]. The genome size (G) of *Cocos*
116 *nucifera* could be estimated by the following formula:

$$117 \quad G = N \times (L - K + 1) / K_depth$$

118 where N represents the total of number of reads, L represents the read length, K represents the k-mer
119 value used in the analysis and K_depth refers to the main peak in the k-mer distribution curve. In our

1 120 calculations, N was 2,049,520,223, L was 100 and K_depth was 71 for K=17. As a result, *Cocos*
2 121 *nucifera* genome was estimated to be 2.42 gigabases (Gb). K-mer size distribution analysis (Figure 2)
3
4 122 indicated that *Cocos nucifera* was a diploid species with low heterozygosity and a high proportion of
5
6 123 repetitive sequences.
7

8 124 We then assembled the *Cocos nucifera* genome using the software SOAPdenovo2
9
10 125 (SOAPdenovo2, RRID:SCR_014986) in three steps: contig construction, scaffold construction and
11
12 126 gap filling. In the contig construction step: the SOAPdenovo2 was run with the parameters ‘pregraph
13
14 127 -K 63 -R -d 1’ to construct de Bruijn graphs from paired-end libraries with insert sizes ranging from
15
16 128 170 to 800 bp. The k-mers from the de Bruijn graphs were then used to form contiguous sequences
17
18 129 (contigs) with the parameters ‘contig -R’ by clipping tips, merging bubbles and removing low
19
20 130 coverage links. In the scaffold construction step: the orders of the contigs were determined by using
21
22 131 paired-end and mate-pair information with parameters ‘map -k 43’ and ‘scaff -F -u’. In more detail,
23
24 132 SOAPdenovo2 maps the reads from paired-end and mate pair libraries to contigs based on a hash
25
26 133 table (keys are unique k-mers on contigs; values are positions). In such cases, two contigs are
27
28 134 considered to be linked if the bridging of the contigs are supported by five paired-end read pairs or
29
30 135 three mate-pair read pairs. In the gap filling step: gaps within scaffolds were filled by utilizing KGF
31
32 136 [7] v1.06 and GapCloser v1.12-r6 (GapCloser, RRID:SCR_015026) [7] with paired-end libraries
33
34 137 (having an insert size from 170 to 800 bp in cases, where one end could be mapped to one contig and
35
36 138 the other end extended into a gap). To optimize the assembled sequence, Rabbit (a Poisson-based
37
38 139 k-mer model software [8]) was used to remove the redundant sequences. A final length of 2.20 Gb for
39
40 140 the scaffolds was obtained and used for further analysis, accounting for 90.91% of the predicted
41
42 141 genome size and larger than the African oil palm and datepalm genomes (Table 2). Meanwhile, the
43
44 142 N50 of the obtained contigs was 72.64 Kb and 418.06 Kb for the scaffolds which have excluding
45
46 143 scaffolds less than 100 bp. The comparison of N50 values for the assembled coconut genome and for
47
48 144 four previously published palm genomes *Elaeis guineensis* [3], *Elaeis oleifera* [3], *Phoenix*
49
50 145 *dactylifera* (PDK30) [4] and *Phoenix dactylifera*(DPV01) [5] is listed in Table 2.
51

52 146 **Genome evaluation**

53
54 147 The 57,304 unigenes (transcript obtained from three different tissues, spear leaves, young leaves and
55
56 148 fruit flesh) as previously reported by Fan et al. [9] were aligned to the assembled genome of *Cocos*
57
58 149 *nucifera* using BLAT (BLAT, RRID:SCR_011919) [10] with default parameters. The alignment
59
60
61
62
63
64
65

150 results indicated that the assembled genome of *Cocos nucifera* covered 96.78% of the expressed
151 unigenes, suggesting a high level of coverage has been reached for the assembled genome (Table 3).

152 We also evaluated the level of genome completeness for the assembled sequences by using
153 BUSCO v2.0 (BUSCO, RRID:SCR_015008) [11], which quantitatively assesses genome
154 completeness using evolutionarily-informed expectations of gene content from near-universal
155 single-copy orthologs selected from OrthoDB v9 (OrthoDB, RRID:SCR_011980;
156 <http://busco.ezlab.org/>, plant set). BUSCO analysis showed that there are separate 90.8% and 3.4% of
157 the 1,440 expected plant genes were identified as complete and fragmented genes respectively, while
158 5.8% of genes were considered to be missing from the assembled coconut genome sequence. The
159 comparative results of the BUSCO estimation in coconut and in the four other palm genome
160 sequences indicates that the smallest fraction of missing genes as predicted by BUSCO was found in
161 the coconut genome assembly (Table 4).

162 **Repeat annotation**

163 We combined homology - based annotation and *de novo* method to identify transposable elements
164 (TEs) and the tandem repeats in the *Cocos nucifera* genome. In homology - based annotation step:
165 TEs were identified by searching against the Repbase library (version 20.04) [12] with RepeatMasker
166 (RepeatMasker, RRID:SCR_012954) (v4.0.5) [13] and RepeatProteinMasker (v4.0.5) [13]. In the *de*
167 *novo* step: *de novo* libraries were constructed based on the genome sequences using the *de novo*
168 prediction program RepeatModeler (RepeatModeler, RRID:SCR_015027) and LTR_FINDER
169 (LTR_FINDER, RRID:SCR_015247) [14] by removing contaminant and multi-copy genes.
170 Subsequently, novel transposable elements were identified and classified using RepeatMasker.
171 Tandem repeat sequences were identified by TRF (Tandem Repeat Finder) software [15] with the
172 following parameters ‘Match = 2, Mismatch = 7, Delta = 7, PM = 80, PI = 10, Minscore = 50 and
173 MaxPeriod = 2000’. The total length of the tandem repeat sequences predicted by the software was
174 151,229,585 bp, comprising 6.86% of the coconut genome. Finally, 1.6 Gb of non-redundant
175 repetitive elements were identified, accounting for 74.48% of the coconut genome. Transposable
176 elements took up 72.75% of the total 1.6Gb of repetitive elements with the long-terminal repeat
177 retrotransposon (LTR) class accounting for 92.23% of all TEs and 67.1% of the coconut genome
178 (Table 5).

179 **Gene prediction**

180 We combined three strategies to predict genes in *Cocos nucifera* genome: homology - based, *de novo*
181 and transcript alignment. For homology - based annotation: the protein sequences of *Arabidopsis*
182 *thaliana* [16], *Oryza sativa*[17], *Sorghum bicolor* [18], *Zea mays* [19], *Elaeis guineensis*, and *Phoenix*
183 *dactylifera* (DPV01) were downloaded from each corresponding source (see “Availability of data
184 sources”). The coconut genome was aligned against these downloaded databases using TBLASTN[20]
185 with parameter ‘-e 1e-5 -F -m 8’ and BLAST results were processed by solar (v0.9) with parameter
186 ‘-aprot 2 genome2 -z’ to determine the candidate gene loci. Next, we extracted the genomic sequences
187 of candidate gene loci along with 1kb flanking sequences, and applied GeneWise 2.2.0 (GeneWise,
188 RRID:SCR_015054) [21] to define the intron - exon boundaries. The genes with pre-stop codon or
189 frame-shifts were excluded from further analysis.

190 For *de novo* prediction: we randomly selected 1000 full-length genes (GeneWise score equal 100,
191 intact structure: start codon, stop codon, perfect intron-exon boundary) from gene models predicted
192 by homology-based methods to train the model parameters for AUGUSTUS 2.5 (Augustus: Gene
193 Prediction, RRID:SCR_008417) [22]. Two software programs, AUGUSTUS 2.5 and GENSCAN
194 (GENSCAN, RRID:SCR_012902) 1.0 [23], were used to do *de novo* prediction on the repeat-masked
195 genome of *Cocos nucifera*. Genes with incomplete structure or protein coding length less than 150bp
196 were filtered out.

197 Subsequently, genes from both homology-based and *de novo* methods were combined to obtain
198 non-redundant gene sets by using GLEAN [24] with the following parameters: minimum coding
199 sequence length 150 bp and maximum intron length 50 kb. Genes were filtered with the same
200 thresholds as were used for homology-based annotation.

201 For transcriptome-based prediction: RNA-seq data (SRR606452) as previously reported by Fan
202 et al. [9] was mapped onto the coconut genome to identify the splice junctions using the software
203 TopHat v2.1.1 (TopHat, RRID:SCR_013035) [25]. The software Cufflinks v2.2.1 (Cufflinks,
204 RRID:SCR_014597) [26] was then used to assemble transcripts with the aligned reads. The coding
205 potential of these transcripts was identified using a fifth-order Hidden Markov Model, which was
206 estimated with the same gene sets used in AUGUSTUS training by train GlimmerHMM, an
207 application in the GlimmerHMM package (GlimmerHMM, RRID:SCR_002654) [27]. The transcripts
208 with intact open reading frames (ORFs) were extracted and the longest transcript was retrieved as a
209 representative of a gene whiles multiple transcripts from on a same locus.

210 Finally, we merged the GLEAN and the transcriptome result to form a comprehensive gene set
211 using an in-house annotation pipeline with the following steps: firstly, all-to-all BLASTP analysis
212 of protein sequences was performed between GLEAN results and transcript assemblies with an
213 E-value cutoff of 1e-10. These transcript assemblies were added to the GLEAN result to form
214 (untranslated region) UTRs or alternative splicing products, depending on whether the coverage
215 and identity of the alignment results reached 0.9 or not. If the transcript assemblies had no
216 BLAST hit with the GLEAN results, these transcript assemblies were added to the final gene set
217 as novel gene. The protocol for integrating GLEAN and transcriptome data is shown in Figure 3.

218 **Gene evaluation**

219 The annotation processes identified 28,039 protein-coding genes (Table 2), which is less than the
220 predicted gene numbers of *Phoenix dactylifera* (PDK30,28,889), *Phoenix dactylifera* (DPV01, 41,660)
221 and *Elaeis guineensis* (34,802). Meanwhile, the BUSCO evaluation showed that 74.1% and 11.2% of
222 1,440 expected plant genes were identified as complete and fragmented, with 14.7% of genes
223 considered missing in the gene sets. The BUSCO results showed that our gene prediction was more
224 complete than that of *Phoenix dactylifera* (PDK30) and *Elaeis guineensis*, but less complete than that
225 of *Phoenix dactylifera* (DPV01) (Table 6),

226 **Gene Function**

227 Gene function annotation was done based on sequence similarity and domains conservation.
228 Firstly, the coconut protein coding genes were aligned against the KEGG (KEGG,
229 RRID:SCR_012773) protein databases [28], SwissProt and TrEMBL [29] using BLASTP at a cut-off
230 E-value threshold of 10^{-5} . Subsequently, the best match from the alignment was used to represent the
231 gene function. We obtained 18,445 KEGG, 18,867 Swissprot and 24,882 Tremble annotated genes.
232 Secondly, InterProScan (InterProScan, RRID:SCR_005829) 5.11-51.0 software [30] was employed to
233 identify the motif and domain based on the public databases Pfam (Pfam, RRID:SCR_004726) [31],
234 PRINTS (PRINTS, RRID:SCR_003412) [32], ProDom (ProDom, RRID:SCR_006969) [33], SMART
235 (SMART, RRID:SCR_005026) [34], PANTHER (PANTHER, RRID:SCR_004869) [35], TIGRFAM
236 (JCVI TIGRFAMS, RRID:SCR_005493) [36] and SUPERFAMILY (SUPERFAMILY,
237 RRID:SCR_007952) [37]. The gene function annotation demonstrated that 21,087 of the coconut
238 proteins had conserved motifs and 1,622 Gene Ontology (GO) terms were assigned to 15,705 coconut
239 proteins from the corresponding InterPro (InterPro, RRID:SCR_006695) entry [38]. In total,

1 240 approximately 89.41% of these genes were functionally annotated using the above methods.

2
3 241 **Gene Family Construction**

4
5 242 Protein sequences of thirteen angiosperms, including *Elaeis guineensis*, *Phoenix dactylifera* (DPV01),
6
7 243 *Sorghum bicolor*, *Prunus persica*, *Solanum tuberosum*, *Glycine max*, *Arabidopsis thaliana*,
8
9 244 *Theobroma cacao*, *Vitis vinifera*, *Musa acuminata*, *Carica papaya*, *Populus trichocarpa* and
10
11 245 *Amborella trichopoda*, were download from each corresponding ftp site (see “Availability of data
12
13 246 sources”). For genes with alternative splicing variants, the longest transcripts were selected to
14
15 247 represent the gene. The gene numbers of *Elaeis guineensis* and *Phoenix dactylifera* (DPV01) were
16
17 248 greatly different from the research paper published in 2013[3, 5], because genes of these two species
18
19 249 were re-predicted using the NCBI Prokaryotic Genome Annotation Pipeline which seemed to be more
20
21 250 reasonable. Similarities between paired sequences were calculated using BLASTP with an E-value
22
23 251 threshold of 1e-5. OrthoMCL (OrthoMCL DB: Ortholog Groups of Protein Sequences,
24
25 252 RRID:SCR_007839) [39] was used to identify gene family based on the similarities of the genes and
26
27 253 a Markov Chain Clustering (MCL) with default parameters. About 79.80% of *Cocos nucifera* genes
28
29 254 were assigned to 14,411 families, of which 282 families only existed in *Cocos nucifera* (coconut
30
31 255 specific families) (Table 7). Figure 4 shows the shared gene families for orthologous genes. There are
32
33 256 544 orthologous families shared by five monocot species and 7706 orthologous families shared by all
34
35 257 monocot and dicot species, suggesting 544 monocot unique functions shared by five monocot species
36
37 258 and 7,706 ancestral functions in the most recent common ancestor of the angiosperms.

38
39
40 259 **Phylogenetic analysis**

41
42
43 260 We extracted 247 single copy orthologous genes derived from the gene family analysis step, and
44
45 261 then aligned the protein sequences of each family with MUSCLE (MUSCLE, RRID:SCR_011812)
46
47 262 (v3.8.31) [40]. Next, the protein alignments were converted to corresponding coding sequences (CDS)
48
49 263 using an in-house Perl script. These coding sequences of each single copy gene family were
50
51 264 concatenated to form one super gene for each species. The nucleotides at position 2 (phase one site)
52
53 265 and 3 (four degenerate site) of codon were extracted separately to construct the phylogenetic tree by
54
55 266 PhyML 3.0 (PhyML, RRID:SCR_014629) [41] using a HKY85 substitution model and a gamma
56
57 267 distribution across sites. The tree constructed by phase one sites was consistent with the tree
58
59 268 constructed by four degenerate sites.

1 269 **Divergence time**

2
3 270 The Bayesian relaxed molecular clock approach was used to estimate species divergence time using
4
5 271 MCMCTREE in PAML (PAML, RRID:SCR_014932) [42], based on the four-degenerate sites and the
6
7 272 data set used in phylogenetic analysis, with previously published calibration times [43] (divergence
8
9 273 between *Arabidopsis thaliana* and *Carica papaya* was 54-90 Mya, divergence between *Arabidopsis*
10
11 274 *thaliana* and *Populus trichocarpa* was 100-120 Mya). The divergence time between coconut and oil
12
13 275 palm is about 46.0 (25.4-83.3) million years ago (Figure 5), which is less than the divergence time
14
15 276 between coconut and date palm.

16
17 277 **Identification of antiporter genes in coconut genome**

18
19 278 Antiporters are transmembrane proteins involved in the exchange of substances within and outside the
20
21 279 membrane. In *Arabidopsis*, the functions of antiporter genes have been well characterized
22
23 280 experimentally, and this gene family was subdivided into thirteen different functional groups. Among
24
25 281 them, three functional clusters involved in Na⁺/H⁺ antiporters, some of which were documented to be
26
27 282 associated with salt tolerance [44, 45].

28
29 283 The amino acid sequences of 70 *antiporter* genes of *Arabidopsis* were downloaded from the
30
31 284 *Arabidopsis* Information Resource TAIR website (TAIR, RRID:SCR_004618;
32
33 285 <http://www.arabidopsis.org>) and used as queries for BLASTP against the predicted proteins in the
34
35 286 *Cocos nucifera* genome with a cut-off e-value of 1e-10. A total of 126 *antiporter* genes were
36
37 287 identified in coconut genome. Using local Hidden Markov Model-based HMMER (v3.0) searches and
38
39 288 the Pfam database, seven antiporter genes were excluded from further analysis because of the lack of
40
41 289 conserved domain. The detailed information of the 119 antiporter genes is listed in Additional file 1.

42
43 290 In order to elucidate the evolutionary relationship and potential functions of the antiporters
44
45 291 identified in the study, we applied phylogenetic analysis of *Arabidopsis* and *C. nucifera* antiporter
46
47 292 proteins using the neighbor joining method (Figure 6). Phylogenetic analysis showed that the 119
48
49 293 antiporter genes from *C. nucifera* can be subdivided into twelve groups and that almost all antiporter
50
51 294 genes were clustered together with the functional groups in *Arabidopsis thaliana*.

52
53 295 Phylogenetic analysis showed that the number of antiporter genes was equal between
54
55 296 *Arabidopsis thaliana* and *C. nucifera* for most groups except for G1 (one of three Na⁺/H⁺ antiporter
56
57 297 family), G3 (carnitine/acylcarnitine translocase family) and G12 (potassium-dependent
58
59

1 298 sodium-calcium exchanger). The G1 group (one of three Na⁺/H⁺ antiporter families) contained only
2 299 one *Arabidopsis* antiporter gene and but 14 *C. nucifera* antiporters (1-At/14-Cn), whereas G3
3 300 (carnitine/acylcarnitine translocase family) contained 1-At/29-Cn, and G13 (Potassium-dependent
4 301 sodium-calcium exchanger) contained 3-At/11-Cn. The Na⁺/H⁺ antiporter family had been reported
5 302 to be associated with salt stress. The expansion of the Na⁺/H⁺ antiporter gene family in coconut palm
6 303 maybe associated with the high salt tolerance of coconut. Meanwhile, carnitine/acylcarnitine
7 304 translocase is involved in fatty acid transport cross the mitochondrial membranes. This gene family
8 305 expansion maybe associated with accumulation of fatty acid in coconut pulp. Moreover, coconut
9 306 water contains a high density of potassium ion, approximately 312 mg potassium ion per 100 g
10 307 coconut water [46]. In this study, the gene number of potassium-dependent sodium-calcium
11 308 exchangers were also detected to be significantly increased compared to *Arabidopsis*.

22 309 **Identification of ion channel genes in coconut genome**

23 310 A total of 67 ion channel genes were identified in the coconut genome (Additional file 2). The amino
24 311 acid sequences of 67 *C. nucifera* and 60 *Arabidopsis* ion channel genes were used to analyze their
25 312 evolutionary relationship (Figure 7). Almost all ion channel genes from *C. nucifera* can be clustered
26 313 into the function groups found in *Arabidopsis thaliana*. The number of ion channel genes was equal
27 314 between *Arabidopsis thaliana* and *Cocos nucifera* in most groups except for G5 (potassium channel).
28 315 Many more genes (21) from *C. nucifera* than from *Arabidopsis thaliana* (9 genes) were present in
29 316 group 5 (potassium channels). The gene family expansion maybe associated with accumulation of
30 317 potassium ions in coconut water.

31 318 **Conclusion**

32 319 *Cocos nucifera* (2n = 32) is an important tropical crop, and is also used as an ornamental plant in
33 320 the tropics. In the present study, we sequenced and *de novo* assembled the coconut genome. A total
34 321 scaffold length of 2.2 Gb was generated, with scaffold N50 of 418 Kb. The divergence time of *Cocos*
35 322 *nucifera* and *Elaeis guineensis* is more recent than that of *Cocos nucifera* and *Phoenix dactylifera*,
36 323 suggesting a closer relationship between *C. nucifera* and *E. guineensis*. Comparative analysis of
37 324 antiporter and ion channels between *C. nucifera* and *Arabidopsis thaliana* showed significant gene
38 325 family expansions maybe involving Na⁺/H⁺ antiporters, carnitine/acylcarnitine translocases,
39 326 potassium-dependent sodium-calcium exchangers, and potassium channels. The expansion of these
40 327 gene families may be associated with adaptation to salt stress, accumulation of fatty acid in coconut

1 328 pulp and potassium ions in coconut water. The data output of the coconut genome will provide a
2 329 valuable resource and reference information for the development of high density molecular makers,
3
4 330 construction of high density linkage maps, detection of QTL (quantitative trait loci), genome-wide
5
6 331 association mapping, and molecular breeding.
7

8 332 **Availability of supporting data**

9
10 333 Supporting data are available in the GigaDB database (GigaDB, RRID:SCR_004002) [47]. Raw data
11
12 334 were deposited in the Sequence Read Archive (SRA539146) with the project accession code
13
14 335 PRJNA374600 for the *Cocos nucifera* genome. Previously published RNA-seq data used for
15
16 336 transcriptome-based prediction is available under accession number SRR606452.
17

18 337 **Availability of other angiosperms data sources**

19
20
21 338 *Arabidopsis thaliana*, *Oryza sativa*, *Sorghum bicolor*, *Zea mays*, *Sorghum bicolor*, *Solanum*
22
23 339 *tuberosum*, *Prunus persica*, *Theobroma cacao*, *Vitis vinifera*, *Musa acuminata*, *Carica papaya*,
24
25 340 *Populus trichocarpa*, *Amborella trichopoda*: <https://phytozome.jgi.doe.gov/pz/portal.html>
26
27 341 (phytozomev9.1)
28

29 342 *Elaeis guineensis*: ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/442/705/GCF_000442705.1_EG5/

30 343 *Phoenix dactylifera* (DPV01):

31
32 344 ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/413/155/GCF_000413155.1_DPV01/

33
34 345 *Phoenix dactylifera* (PDK30):

35
36 346 <http://qatar-weill.cornell.edu/research/datepalmGenome/download.html>
37

38 347 **Competing interests**

39
40
41 348 The authors declare that they have no competing interests.
42

43 349 **Funding**

44
45 350 This study was supported by International Science and Technology Cooperation projects of Hainan
46
47 351 Province (No. KJHZ2014-24), Hainan Natural Science Foundation (No. 313058), the major
48
49 352 Technology Project of Hainan (No. ZDZX2013023-1), the fundamental Scientific Research Funds for
50
51 353 Chinese Academy of Tropical Agriculture Sciences (CATAS-No. 1630032012044, 1630052014002,
52
53 354 1630052015050, 1630152017019, and 1630152016006), Central Public-interest Scientific Institution
54
55 355 Basal Research Fund for Innovative Research Team Program of CATAS (No.17CXTD-28).
56

57 356 **Author's contribution**

1 357 YX, HF, YY, MP, QL, AG designed the study and contribute to the project coordination; XY, PX, WX
2 358 wrote the paper; LZ, JL, YW collected the samples and extracted the genomic DNA; YX, BL, BS, JX,
3
4 359 AA, EI, NL conducted the genome analyses.
5

6 360 **Acknowledgements**

7
8 361 Annaliese S. Mason is gratefully acknowledged for assistance with language editing and manuscript
9
10 362 revisions.
11

12 363

13 364 **References**

- 14
15
16 365 1. Batugal P, V Ramanatha Rao and J Oliver, editors. Coconut Genetic Resources. International
17 366 Plant Genetic Resources Institute – Regional Office for Asia, the Pacific and Oceania
18 367 (IPGRI-APO) Serdang, Selangor DE, Malaysia; 2005.
19
20 368 2. Tang B, Tang M, Chen C, Qiu P, Liu Q, Wang M, et al. Characteristics of soil fauna community
21 369 in the Dongjiao coconut plantation ecosystem in Hainan, China. *Acta Ecologica Sinica*.
22 370 2006;26(1):26-32. doi:[http://dx.doi.org/10.1016/S1872-2032\(06\)60003-6](http://dx.doi.org/10.1016/S1872-2032(06)60003-6).
23
24 371 3. Singh R, Ong-Abdullah M, Low ET, Manaf MA, Rosli R, Nookiah R, et al. Oil palm genome
25 372 sequence reveals divergence of interfertile species in Old and New worlds. *Nature*.
26 373 2013;500(7462):335-9. doi:10.1038/nature12309.
27
28 374 4. Al-Dous EK, George B, Al-Mahmoud ME, Al-Jaber MY, Wang H, Salameh YM, et al. De novo
29 375 genome sequencing and comparative genomics of date palm (*Phoenix dactylifera*). *Nature*
30 376 *biotechnology*. 2011;29(6):521-7. doi:10.1038/nbt.1860.
31
32 377 5. Al-Mssallem IS, Hu S, Zhang X, Lin Q, Liu W, Tan J, et al. Genome sequence of the date palm
33 378 *Phoenix dactylifera* L. *Nature communications*. 2013;4:2274. doi:10.1038/ncomms3274.
34
35 379 6. Murray MG and Thompson WF. Rapid isolation of high molecular weight plant DNA. *Nucleic*
36 380 *acids research*. 1980;8(19):4321-5.
37
38 381 7. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, et al. SOAPdenovo2: an empirically improved
39 382 memory-efficient short-read de novo assembler. *Gigascience*. 2012;1(1):18.
40 383 doi:10.1186/2047-217X-1-18.
41
42 384 8. Zhan, D. Rabbit Genome Assembler. 2017.
43 385 <https://github.com/gigascience/rabbit-genome-assembler>
44
45 386 9. Fan H, Xiao Y, Yang Y, Xia W, Mason AS, Xia Z, et al. RNA-Seq analysis of *Cocos nucifera*:
46 387 transcriptome sequencing and de novo assembly for subsequent functional genomics
47 388 approaches. *PloS one*. 2013;8(3):e59997. doi:10.1371/journal.pone.0059997.
48
49 389 10. Kent WJ. BLAT--the BLAST-like alignment tool. *Genome Research*. 2002;12(4):656-64.
50 390 doi:10.1101/gr.229202. .
51
52 391 11. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV and Zdobnov EM. BUSCO: assessing
53 392 genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*
54 393 (Oxford, England). 2015;31(19):3210-2. doi:10.1093/bioinformatics/btv351.
55
56 394 12. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O and Walichiewicz J. Repbase
57 395 Update, a database of eukaryotic repetitive elements. *Cytogenetic and genome research*.
58 396 2005;110(1-4):462-7. doi:10.1159/000084979.

1 397 13. Tarailo-Graovac M and Chen N. Using RepeatMasker to identify repetitive elements in
2 398 genomic sequences. *Current protocols in bioinformatics*. 2009;Chapter 4:Unit 4.10.
3 399 doi:10.1002/0471250953.bi0410s25.

4 400 14. Xu Z and Wang H. LTR_FINDER: an efficient tool for the prediction of full-length LTR
5 401 retrotransposons. *Nucleic acids research*. 2007;35(Web Server issue):W265-8.
6 402 doi:10.1093/nar/gkm286.

7 403 15. Benson G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic acids*
8 404 *research*. 1999;27(2):573-80.

9 405 16. The Arabidopsis Genome Initiative, Analysis of the genome sequence of the flowering plant
10 406 *Arabidopsis thaliana*. *Nature*. 2000;408(6814):796-815.
11 407 doi:http://www.nature.com/nature/journal/v408/n6814/supinfo/408796a0_S1.html.

12 408 17. Goff SA, Ricke D, Lan TH, Presting G, Wang R, Dunn M, et al. A draft sequence of the rice
13 409 genome (*Oryza sativa* L. ssp. *japonica*). *Science (New York, NY)*. 2002;296(5565):92-100.
14 410 doi:10.1126/science.1068275.

15 411 18. Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, et al. The
16 412 *Sorghum bicolor* genome and the diversification of grasses. *Nature*. 2009;457(7229):551-6.
17 413 doi:10.1038/nature07723.

18 414 19. Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, et al. The B73 maize genome:
19 415 complexity, diversity, and dynamics. *Science (New York, NY)*. 2009;326(5956):1112-5.
20 416 doi:10.1126/science.1178534.

21 417 20. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and
22 418 PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*.
23 419 1997;25(17):3389-402.

24 420 21. Birney E, Clamp M and Durbin R. GeneWise and Genomewise. *Genome Research*.
25 421 2004;14(5):988-95. doi:10.1101/gr.1865504.

26 422 22. Stanke M, Keller O, Gunduz I, Hayes A, Waack S and Morgenstern B. AUGUSTUS: ab initio
27 423 prediction of alternative transcripts. *Nucleic acids research*. 2006;34(Web Server
28 424 issue):W435-9. doi:10.1093/nar/gkl200.

29 425 23. Burge C and Karlin S. Prediction of complete gene structures in human genomic DNA.
30 426 *Journal of molecular biology*. 1997;268(1):78-94. doi:10.1006/jmbi.1997.0951.

31 427 24. Elsik CG, Mackey AJ, Reese JT, Milshina NV, Roos DS and Weinstock GM. Creating a honey
32 428 bee consensus gene set. *Genome biology*. 2007;8(1):R13. doi:10.1186/gb-2007-8-1-r13.

33 429 25. Trapnell C, Pachter L and Salzberg SL. TopHat: discovering splice junctions with RNA-Seq.
34 430 *Bioinformatics (Oxford, England)*. 2009;25(9):1105-11. doi:10.1093/bioinformatics/btp120.

35 431 26. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript
36 432 assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform
37 433 switching during cell differentiation. *Nature biotechnology*. 2010;28(5):511-5.
38 434 doi:10.1038/nbt.1621.

39 435 27. Majoros WH, Pertea M and Salzberg SL. TigrScan and GlimmerHMM: two open source ab
40 436 initio eukaryotic gene-finders. *Bioinformatics (Oxford, England)*. 2004;20(16):2878-9.
41 437 doi:10.1093/bioinformatics/bth315.

42 438 28. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H and Kanehisa M. KEGG: Kyoto Encyclopedia of
43 439 Genes and Genomes. *Nucleic acids research*. 1999;27(1):29-34.

44 440 29. Bairoch A and Apweiler R. The SWISS-PROT protein sequence database and its supplement

441 TrEMBL in 2000. *Nucleic acids research*. 2000;28(1):45-8.

1 442 30. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale
2 443 protein function classification. *Bioinformatics* (Oxford, England). 2014;30(9):1236-40.
3 444 doi:10.1093/bioinformatics/btu031.

4 445 31. Bateman A, Birney E, Durbin R, Eddy SR, Howe KL and Sonnhammer EL. The Pfam protein
5 446 families database. *Nucleic acids research*. 2000;28(1):263-6.

6 447 32. Attwood TK, Croning MDR, Flower DR, Lewis AP, Mabey JE, Scordis P, et al. PRINTS-S: the
7 448 database formerly known as PRINTS. *Nucleic acids research*. 2000;28(1):225-7.

8 449 33. Corpet F, Gouzy J and Kahn D. Recent improvements of the ProDom database of protein
9 450 domain families. *Nucleic acids research*. 1999;27(1):263-7.

10 451 34. Schultz J, Copley RR, Doerks T, Ponting CP and Bork P. SMART: a web-based tool for the
11 452 study of genetically mobile domains. *Nucleic acids research*. 2000;28(1):231-4.

12 453 35. Mi H, Huang X, Muruganujan A, Tang H, Mills C, Kang D, et al. PANTHER version 11:
13 454 expanded annotation data from Gene Ontology and Reactome pathways, and data analysis
14 455 tool enhancements. *Nucleic acids research*. 2017;45(Database issue):D183-D9.
15 456 doi:10.1093/nar/gkw1138.

16 457 36. Selengut JD, Haft DH, Davidsen T, Ganapathy A, Gwinn-Giglio M, Nelson WC, et al.
17 458 TIGRFAMs and Genome Properties: tools for the assignment of molecular function and
18 459 biological process in prokaryotic genomes. *Nucleic acids research*. 2007;35(Database
19 460 issue):D260-D4. doi:10.1093/nar/gkl1043.

20 461 37. Wilson D, Pethica R, Zhou Y, Talbot C, Vogel C, Madera M, et al. SUPERFAMILY--sophisticated
21 462 comparative genomics, data mining, visualization and phylogeny. *Nucleic acids research*.
22 463 2009;37(Database issue):D380-6. doi:10.1093/nar/gkn762.

23 464 38. Burge S, Kelly E, Lonsdale D, Mutowo-Muellenet P, McAnulla C, Mitchell A, et al. Manual GO
24 465 annotation of predictive protein signatures: the InterPro approach to GO curation. *Database*
25 466 : the journal of biological databases and curation. 2012;2012:bar068.
26 467 doi:10.1093/database/bar068.

27 468 39. Li L, Stoeckert CJ, Jr. and Roos DS. OrthoMCL: identification of ortholog groups for
28 469 eukaryotic genomes. *Genome Research*. 2003;13(9):2178-89. doi:10.1101/gr.1224503.

29 470 40. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput.
30 471 *Nucleic acids research*. 2004;32(5):1792-7. doi:10.1093/nar/gkh340.

31 472 41. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W and Gascuel O. New algorithms
32 473 and methods to estimate maximum-likelihood phylogenies: assessing the performance of
33 474 PhyML 3.0. *Systematic biology*. 2010;59(3):307-21. doi:10.1093/sysbio/syq010.

34 475 42. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular biology and*
35 476 *evolution*. 2007;24(8):1586-91. doi:10.1093/molbev/msm088.

36 477 43. Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, et al. The genome of
37 478 black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* (New York, NY).
38 479 2006;313(5793):1596-604. doi:10.1126/science.1128691.

39 480 44. Shi H, Lee BH, Wu SJ and Zhu JK. Overexpression of a plasma membrane Na⁺/H⁺ antiporter
40 481 gene improves salt tolerance in *Arabidopsis thaliana*. *Nature biotechnology*.
41 482 2003;21(1):81-5. doi:10.1038/nbt766.

42 483 45. Brini F, Hanin M, Mezghani I, Berkowitz GA and Masmoudi K. Overexpression of wheat
43 484 Na⁺/H⁺ antiporter TNHX1 and H⁺-pyrophosphatase TVP1 improve salt- and drought-stress

1 485 tolerance in *Arabidopsis thaliana* plants. *Journal of experimental botany*. 2007;58(2):301-8.
2 486 doi:10.1093/jxb/erl251.

3 487 46. Yong JW, Ge L, Ng YF and Tan SN. The chemical composition and biological properties of
4 488 coconut (*Cocos nucifera* L.) water. *Molecules* (Basel, Switzerland). 2009;14(12):5144-64.
5 489 doi:10.3390/molecules14125144.

6
7 490 47. Xiao Y, Xu P, Fan H, Baudouin L, Xia W, Bocs S et al. Supporting data for "The genome draft
8 491 of coconut (*Cocos nucifera*)". *Gigascience Database*. 2017. <http://dx.doi.org/10.5524/100347>
9 492
10 493
11 494
12 495
13 496
14 497
15 498
16 499
17 500
18 501
19 502
20 503
21 504
22 505
23 506
24 507
25 508
26 509
27 510
28 511
29 512
30 513
31 514
32 515
33 516
34 517
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

1 518

2 519

3 520

4 521

5 522

6 523 **Tables**

7 524 Table 1 Data outputs produced by sequencing different insert size libraries

Library type	Lane	Reads Length(bp)	Insert Size(bp)	Raw data (Gb)	Clean data(Gb)
PE101	3	100	170	128.75(53.20)	111.32(46)
PE251	2	250	450	73.86(30.52)	56.42(23.31)
PE101	2	100	500	64(26.45)	65.11(26.90)
PE101	2	100	800	78.16(32.30)	64.90(26.82)
MP50	3	49	2000	128.6(53.14)	60.70(25.08)
MP50	2	49	5000	71.75(29.65)	18.62(7.69)
MP50	2	49	10000	74.65(30.85)	18.53(7.66)
MP50	2	49	20000	70.7(29.21)	19.35(7.99)
MP50	1	49	40000	24.2(10.08)	4.13(1.71)
Total	19			714.67(295.32)	419.08(173.17)

8 525 Note: The sequencing depth was shown in parentheses, calculated based on a genome size of 2.42G. Clean data
9 526 were obtained by filtering raw data with low-quality and duplicate reads. PE: paired-end, MP: mate pair.

10 527

11 528 Table 2 Comparison analysis of genome sizes, assembly and annotation of four palmae species, including

12 529 coconut, *Phoenix dactylifera* (PDK30 and DPV01, two different versions), *Elaeis guineensis* (EG), and *Elaeis*
13 530 *oleifera* (EO)

Species	Sequencing technology	Sequence coverage	Estimated size(Gb)	Assembly size(Gb)	Contig N50(Kb)	Scaffold N50(Kb)	Gene Number	TEs percent (%)
<i>Phoenix dactylifera</i> (PDK30)	Illumina GAIIx	53.4x	0.66	0.38	6.44	30.48	28,889	23.6
<i>Phoenix dactylifera</i> (DPV01)	454,SOLiD, ABI3730	139x	0.67	0.56	10.81	334.08	41,660	38.87
<i>Elaeis guineensis</i> (African oil palm)	454	16X	1.8	1.54	9.37	1045.41	34,802	43.24
<i>Elaeis oleifera</i> (American oil palm)	454	16x	1.8	1.40	8.45	333.11	--	--
<i>Cocos nucifera</i> (Hai nan Tall)	Illumina HiSeq	173X	2.42	2.20	72.64	418.07	28,039	72.75

14 531 Note: Coconut: *Cocos nucifera* (Hainan Tall); PDK30: *Phoenix dactylifera* (PDK30); DPV01: *Phoenix*

532 *dactylifera* (DPV01); EG: *Elaeis guineensis* (African oil palm E5 build); EO *Elaeis oleifera* (American oil palm,
 533 O8-build); TE results were obtained using the same pipeline as for the coconut genome

534
 535
 536
 537
 538
 539
 540

541 Table 3 The gene coverage of *Cocos nucifera* based on transcriptome data

Dataset	Number	Total length (bp)	Base coverage by assembly	Sequence coverage by assembly (%)
All	57,304	43,090,665	96.78	99.57
>200bp	57,304	43,090,665	96.78	99.57
>500bp	25,713	33,470,388	96.36	99.85
>1000bp	13,796	25,004,919	95.99	99.94

542

543 Table 4 The comparative analysis of assembly results of five palm species with BUSCO software, including
 544 coconut, *Phoenix dactylifera* (PDK30 and DPV01, two varieties), *Elaeis guineensis* (EG), and *Elaeis oleifera*
 545 (EO)

BUSCOs	Coconut		PDK30		DPV01		EG		EO	
	N	P (%)	N	P (%)	N	P (%)	N	P (%)	N	P (%)
Total	1440		1440		1440		1440		1440	
Complete single-copy	1192	82.8	1042	72.4	1160	80.6	1100	76.4	1004	69.7
Complete duplicated	115	8.0	81	5.6	134	9.3	116	8.1	63	4.4
Fragment	49	3.4	98	6.8	42	2.9	60	4.2	84	5.8
Missing	84	5.8	219	15.2	104	7.2	164	11.3	289	20.1

546 Note: Coconut: *Cocos nucifera* (the Hainan Tall); PDK30: *Phoenix dactylifera* (PDK30); DPV01: *Phoenix*
 547 *dactylifera* (DPV01); EG: *Elaeis guineensis* (African oil palm E5 build); EO *Elaeis oleifera* (American oil palm,
 548 O8-build);

549

550 Table 5 Classification of predicted transposable elements in the coconut genome

	Repabse TEs length	Protein TEs length	<i>De novo</i> TEs length	Combined TEs length	percentage
DNA	20,936,158	24,655,089	35,131,002	58,119,982	2.64
LINE	4,251,185	9,631,472	7,610,172	19,197,064	0.87
SINE	85,717	0.00	186,364	270,055	0.012
LTR	361,968,154	512,700,933	1,419,281,798	1,478,182,089	67.10
Other	8,145	0.00	0.00	8,145	0.0004
Unknown	0.00	12,360	139,084,335	139,096,695	6.31

1
 2
 3
 4
 5
 6
 7
 8
 9
 10
 11
 12
 13
 14
 15
 16
 17
 18
 19
 20
 21
 22
 23
 24
 25
 26
 27
 28
 29
 30
 31
 32
 33
 34
 35
 36
 37
 38
 39
 40
 41
 42
 43
 44
 45
 46
 47
 48
 49
 50
 51
 52
 53
 54
 55
 56
 57
 58
 59
 60
 61
 62
 63
 64
 65

Total	385,037,442	546,965,774	1,552,582,881	1,602,630,396	72.75
-------	-------------	-------------	---------------	---------------	-------

551 Note: Repbase TEs means RepeatMask against Repbase; Protein TEs means RepeatProteinMask result against
 552 Repbase protein; *De novo* TEs means RepeatMask against the *de novo* library; Combined TEs: the combined
 553 results of these three steps.

554

555

556

557 Table 6 The comparative analysis of gene prediction results of four palm species with BUSCO software

BUSCOs	Coconut		PDK30		DPV01		EG	
	N	P (%)	N	P (%)	N	P (%)	N	P (%)
Total	1440		1440		1440		1440	
Complete single-copy	965	74.1	748	51.9	1195	83.0	555	38.5
Complete duplicated	102	7.1	81	5.6	159	11.0	53	3.7
Fragment	162	11.2	255	17.7	44	3.1	270	18.8
Missing	211	14.7	356	24.8	42	2.9	562	39.0

558 Note: Coconut: *Cocos nucifera* (the Hainan Tall); PDK30: *Phoenix dactylifera* (PDK30); DPV01: *Phoenix*
 559 *dactylifera* (DPV01); EG: *Elaeis guineensis* (African oil palm E5 build); The gene of *Elaeis oleifera* (American
 560 oil palm, O8-build) was missing, not attained from the public database;

561

562 Table 7 Statistical analysis of gene families of different species

Species	Genes number	Genes in families	Unclustered genes	Family number	Unique families	Average genes per family
<i>C. nucifera</i>	28,039	22,376	5,663	14,411	282	1.55
<i>E. guineensis</i>	30,430	22,021	8,409	13,415	262	1.64
<i>P. dactylifera</i>	24,908	22,193	2,715	14,074	112	1.58
<i>S. bicolor</i>	27,159	22,016	5,143	12,992	916	1.69
<i>P. persica</i>	27,792	24,276	3,516	14,443	497	1.68
<i>S. tuberosum</i>	34,879	28,288	6,591	13,206	1,119	2.14
<i>G. max</i>	42,859	38,104	4,755	14,589	1,145	2.61
<i>A. thaliana</i>	26,637	22,990	3,647	13,292	674	1.73
<i>T. cacao</i>	28,624	23,776	4,848	14,928	625	1.59
<i>V. vinifera</i>	25,329	19,122	6,207	13,309	599	1.44
<i>M. acuminata</i>	36,538	24,354	12,184	13,089	620	1.86

563

564

565

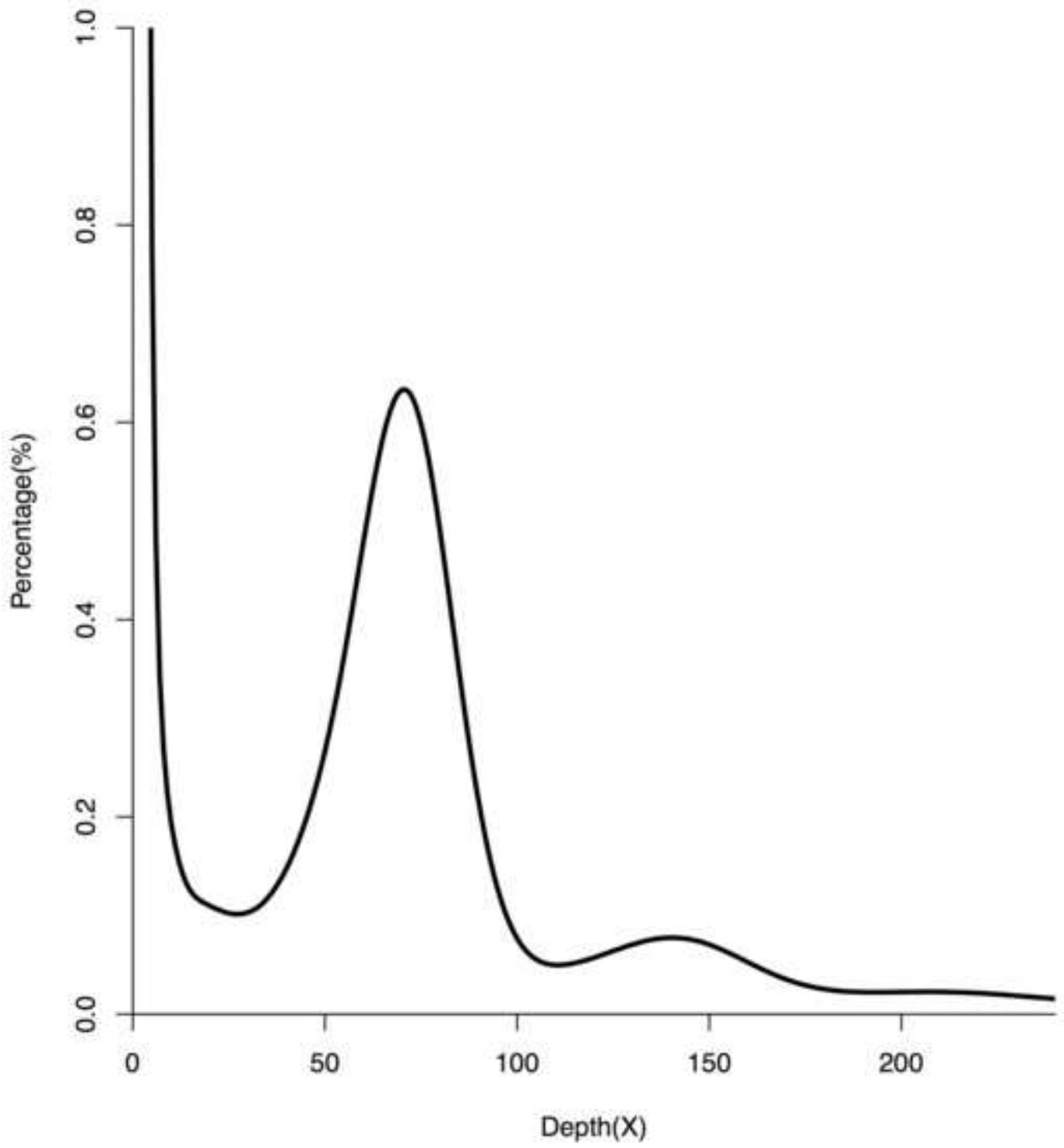
566

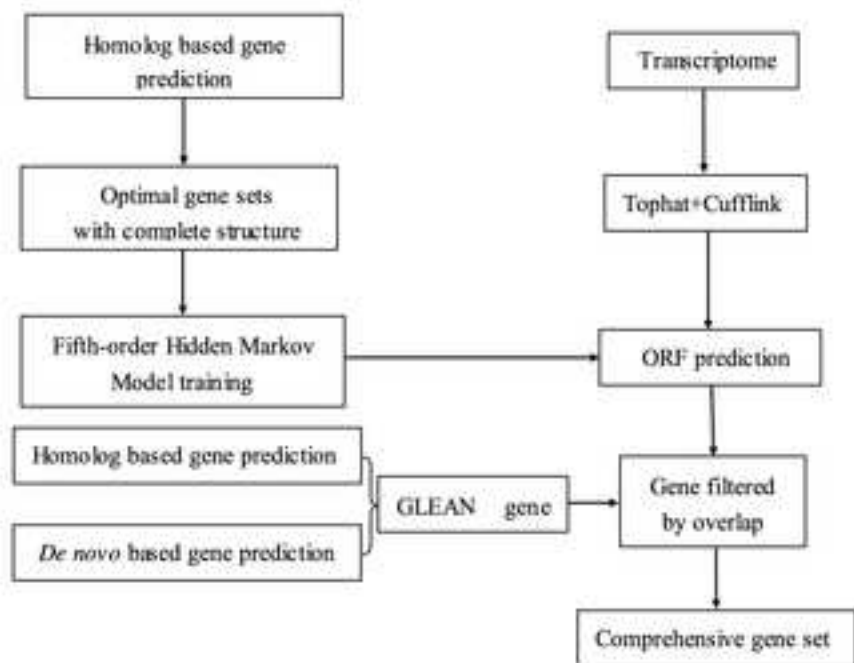
567

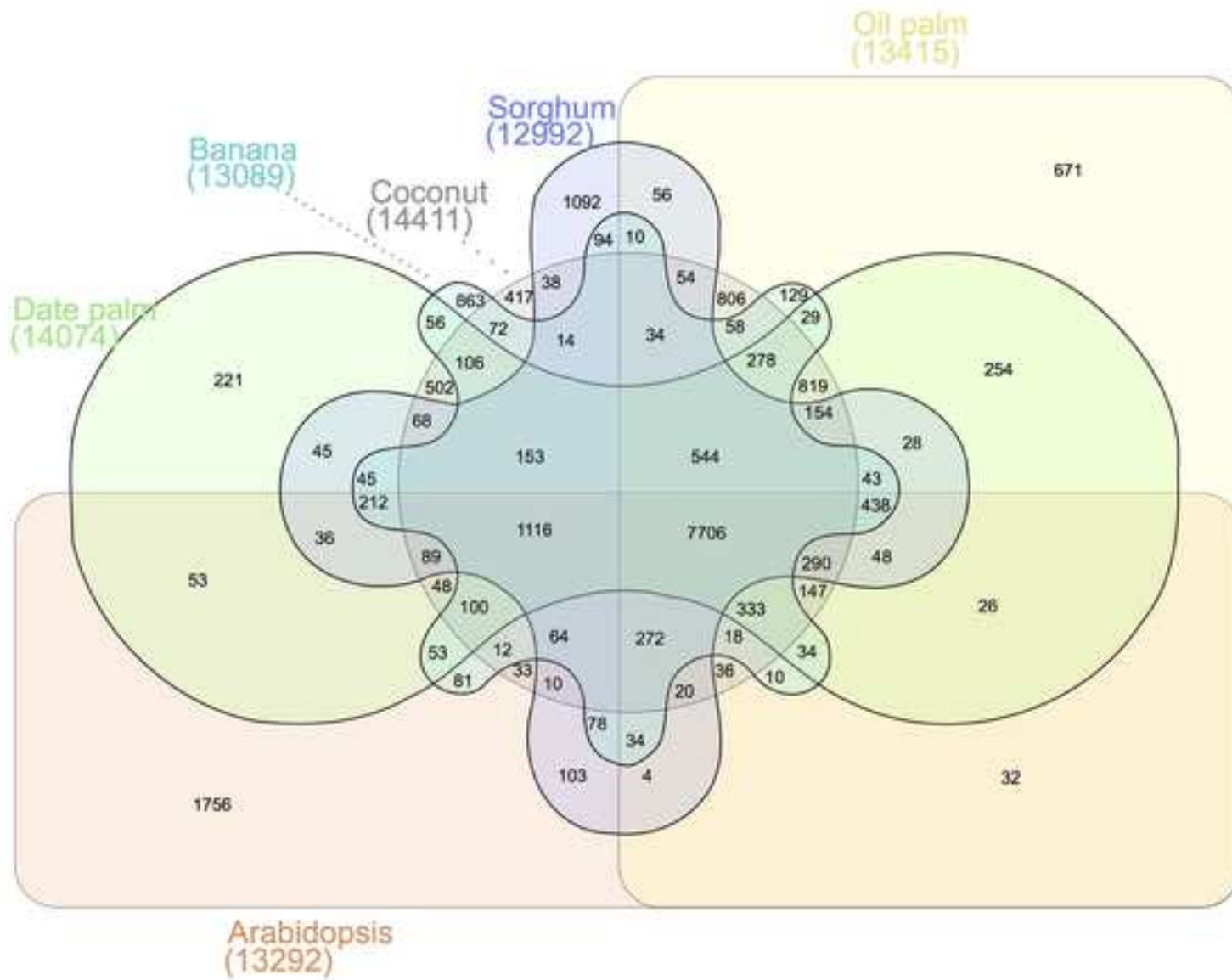
568

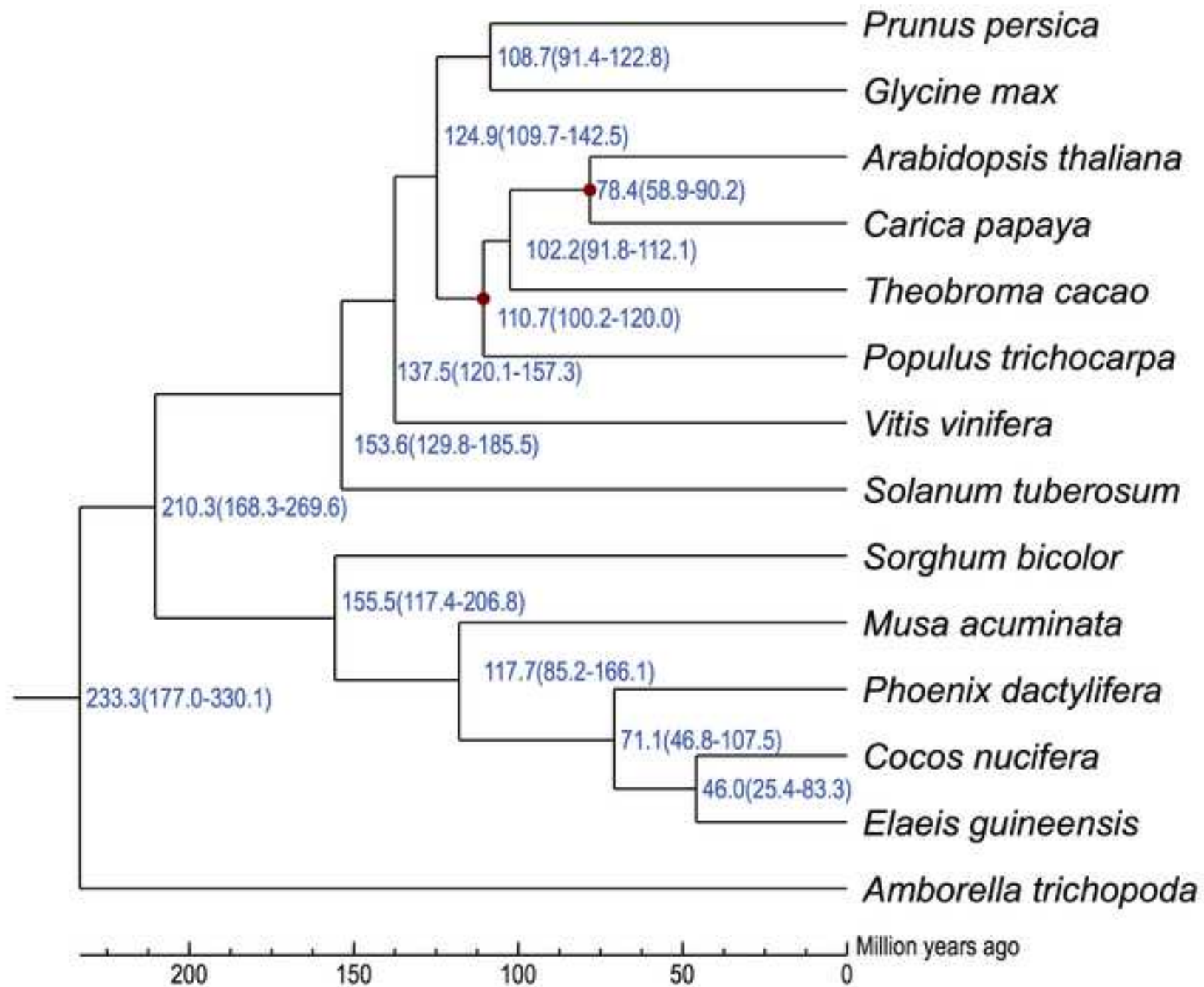
1 569
2 570
3
4 571
5
6 572
7
8 573 **Figure legends**
9
10 574 Figure 1 Morphological characteristic of coconut tree (A), spica (B), female flower (C), Male flower
11
12 575 (D), coconut nut (E), coconut nut without skin (F), and vertical section of coconut nut (G).
13
14 576 Figure 2 Kmer analysis of the coconut genome.
15
16 577 Figure 3 The protocol for integrating GLEAN and transcriptome data.
17
18 578 Figure 4 Groups of orthologues shared among the angiosperms *Cocos nucifera* (Coconut), *Elaeis*
19
20 579 *guineensis* (Oil palm), *Phoenix dactylifera* (Date palm), *Sorghum bicolor* (Sorghum), *Musa*
21
22 580 *acuminata* (Banana) and *Arabidopsis thaliana* (Arabidopsis). Venn diagram generated by
23
24 581 <http://www.interactivenn.net/>.
25
26 582 Figure 5. Estimation of divergence time. The blue numbers on the nodes are the divergence time from
27
28 583 present (million years ago, Mya), the red nodes indicated the previously published calibration times.
29
30 584 Figure 6. Phylogenetic tree of antiporter genes from *C. nucifera* and *Arabidopsis thaliana*. Every
31
32 585 cluster is indicated with a different colored arc line arc. The potential function of every cluster is
33
34 586 indicated with the function groups found in *Arabidopsis thaliana*. Colored stars indicate antiporter
35
36 587 genes of *C. nucifera*.
37
38 588 Figure 7. Phylogenetic tree of ion channel genes from *C. nucifera* and *Arabidopsis thaliana*. Every
39
40 589 cluster was indicated with different colored arc line arc. The potential function of every cluster was
41
42 590 indicated with the function groups found in *Arabidopsis thaliana*. Colored stars indicate ion channel
43
44 591 genes of *C. nucifera*.
45
46
47
48
49
50
51
52
53
54
55
56
57
58 592
59
60 593

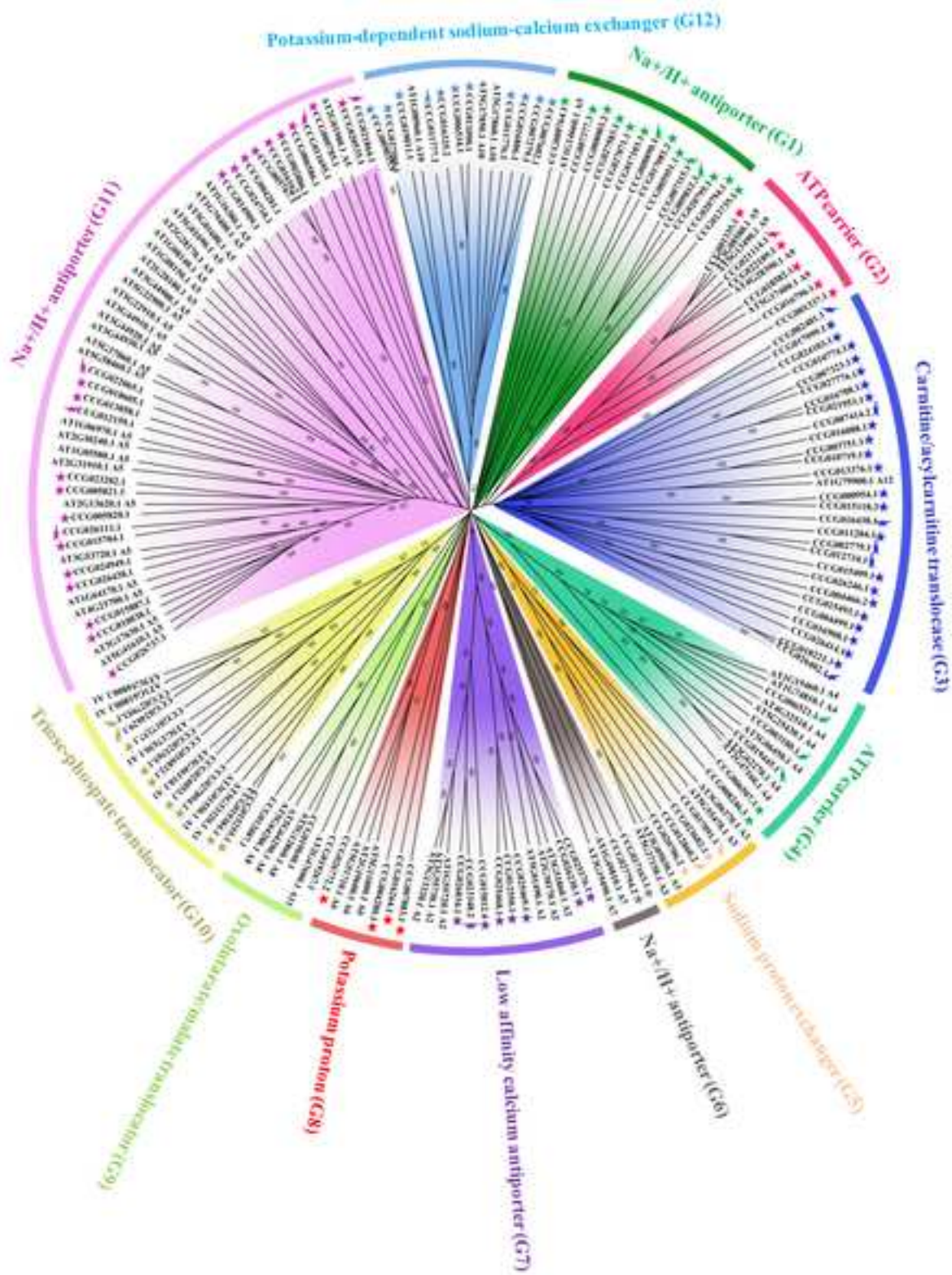
1	594	
2	595	
3		
4	596	
5		
6	597	
7		
8	598	Additional files
9		
10	599	
11		
12	600	Additional file 1 Identification and characterization of antiporter genes in the genome of Cocos
13		
14	601	nucifera
15		
16	602	
17		
18	603	Additional file 2 Identification and characterization of ion channel genes in the genome of Cocos
19		
20	604	nucifera
21		
22		
23		
24		
25		
26		
27		
28		
29		
30		
31		
32		
33		
34		
35		
36		
37		
38		
39		
40		
41		
42		
43		
44		
45		
46		
47		
48		
49		
50		
51		
52		
53		
54		
55		
56		
57		
58		
59		
60		
61		
62		
63		
64		
65		

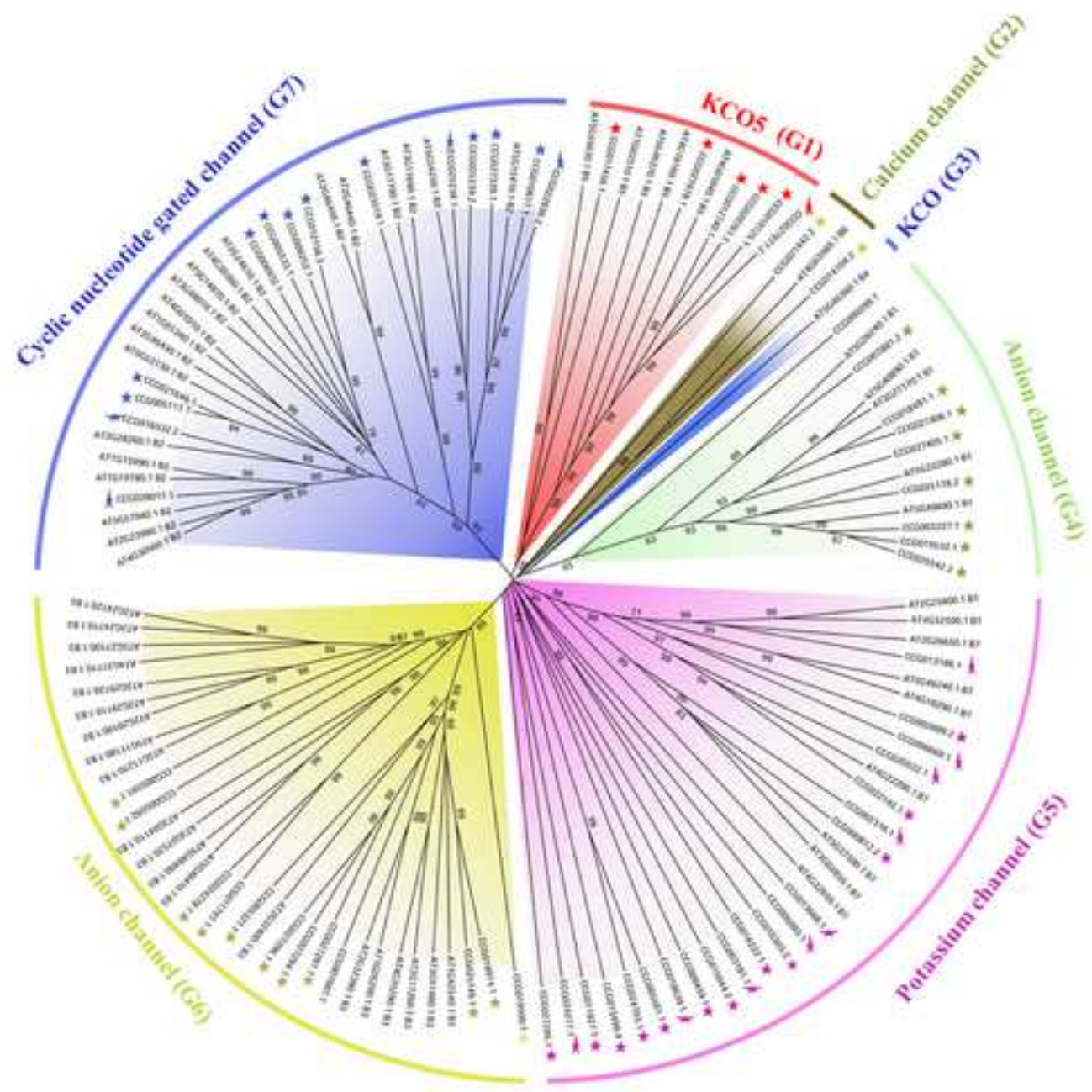




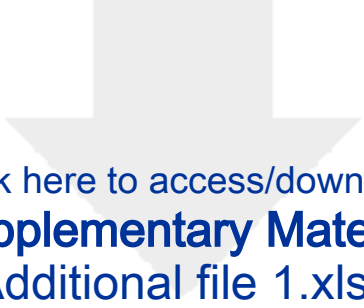







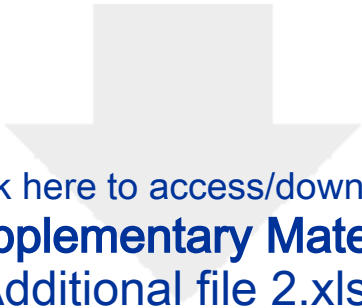






Click here to access/download
Supplementary Material
Additional file 1.xlsx





Click here to access/download
Supplementary Material
Additional file 2.xlsx

