

# GigaScience

## The genome draft of the Coconut (*Cocos nucifera*)

--Manuscript Draft--

<b>Manuscript Number:</b>	GIGA-D-17-00038R1	
<b>Full Title:</b>	The genome draft of the Coconut ( <i>Cocos nucifera</i> )	
<b>Article Type:</b>	Data Note	
<b>Funding Information:</b>	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
<b>Abstract:</b>	<p><b>Background</b> Coconut palm (<i>Cocos nucifera</i>, <math>2n = 32</math>), 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 (<a href="http://www.fao.org/faostat/en/">www.fao.org/faostat/en/</a>). 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><b>Findings</b> 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× 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<sup>+</sup>/H<sup>+</sup> 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>
<b>Corresponding Author:</b>	Yaodong Yang, Ph.D Coconut Research Institute Wenchang City, Hainan CHINA
<b>Corresponding Author Secondary Information:</b>	
<b>Corresponding Author's Institution:</b>	Coconut Research Institute
<b>Corresponding Author's Secondary Institution:</b>	
<b>First Author:</b>	Yong Xiao
<b>First Author Secondary Information:</b>	
<b>Order of Authors:</b>	<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>
<b>Order of Authors Secondary Information:</b>	
<b>Response to Reviewers:</b>	<p>Response to editor and reviewers</p> <p>Dear editor and reviewers</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 a thorough revision to the ms based on all of comments from editor and reviewers. Each</p>

comments raised by the reviewers had been carefully answered in the response sheet. We hope the revised version can meet the requirement of "GigaScience"

Sincerely yours,

Yaodong Yang

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

Response to editor and reviewers,

Reviewer 1

1.Line 40: in 93 countries -> the introduction (line 70) say 89 countries

>>>Response: Thank you for your suggestion. We have re-checked the document reported by Batugal et al., 2005. The corresponding revision has been done in the Introduction part of the revised manuscript.

2.11 million ha ->the introduction (line 72) says 12 million ha

>>>Response: Thank you for your suggestion; we have re-checked the plant area of coconut in the website of Food and Agriculture Organization of the United Nations (<http://www.fao.org/faostat/en/>). The corresponding revision has been done in the Abstract part of the revised manuscript.

3.Hinders progress in genetic breeding. Do you mean 'marker assisted breeding' or 'genomic assisted breeding'?

>>>Response: Thank you for your suggestion; we meant to say 'conventional breeding'. Revisions have been made in the Abstract part of the revised manuscript to make our opinions clearer.

4.Genetic improvement is slow. Do you mean trait improvement with marker or genetic assisted

>>>Response: We meant to say the improvement made by 'conventional breeding' is slow. The corresponding revision has been done in the revised manuscript.

5.Line 48: The coverage does not add up. 714.67 Gb on a 2.42 Gb genome is 295× coverage. In any case, only the coverage of the cleaned reads should be shown (177×)

>>>Response: Thank you for your suggestion; in revised manuscript, only the cleaned reads were used for the coverage depth analysis and the coverage is 173.32× read depth.

6.Line54: Do you mean 41,166 genes

>>>Response: Thank you for your suggestion; we have re-checked the annotated gene number for datepalm based on the document reported by Al-Mssallem et al., 2013 and 41 660 genes were annotated. The corresponding revisions have been made in the Abstract part of revised manuscript.

7.Line60: space missing between facilitating and future

>>>Response: Thank you for your suggestion, a space has been added between facilitating and future.

8.Line 61: should be 'molecular assisted breeding'

>>>Response: Thank you for your suggestion, corresponding revisions have been done in the Abstract part of revised version.

9.Line 78: '...wide range to environment...' -> unclear, should be explained. Also 'environment'

>>>Response: Some sentences have been added to the revised manuscript for explaining '...wide range to environment...' in Line 240– Line 242|Page 3.

10.Line78: '...especially for high tolerance to high salt density.', please clarify

>>>Response: Coconut palm can disseminate through ocean currents: floating nuts sprout and grow naturally upon washing up on beaches. The ability to adapt to a high salt environment is closely related to this dissemination feature and to these natural growth conditions. Corresponding revision has been done in Line 243– Line 244|Page 3 of revised manuscript.

11.Line 80: '...making it possible to understand its adaptation to high salinity.' You do not investigate this, you should change the statement to something milder such as: 'This study forms the basis for future research investigating the coconuts tolerance to salt stress'

>>>Response: Thank you for your suggestion, We also present the genome sequence of HAT coconut and added an analysis of the antiporter and ion channel gene families, relevant to salinity tolerance, into the revised version. Corresponding revision had been added into in Line 237– Line 238|Page 3.

12.Line 82: provide references. The way this sentence reads at the moment, make it seem like you are also reporting those genome sequence.

>>>Response: The corresponding references have been added into Line 423|Page 4 of revised manuscript.

13.Line 92: space between 'Illumina', 'Hiseq2000' and 'sequencer'

>>>Response: Two spaces had been added into between Illumina, Hiseq2000 and sequencer in Line 436|Page 4 of revised manuscript.

14.Line129: The data shows that you have higher coverage and a longer N50, it does not show that the assembly is of better quality.

>>>Response: Thank you for your suggestion, the sentence has been replace by other sentence: "The comparative results of the BUSCO estimation in coconut and in the four other palm genome sequences indicates that the smallest fraction of missing genes as predicted by BUSCO was found in the coconut genome assmebly", in Line 724 – Line 726|Page 6 of revised version.

15.Line 131: 'tissues', not 'issues'

>>>Response: Thank you for your suggestion, corresponding revisions has been done in the revised version.

16.Line134: table 4 and 5 are mixed up

>>>Response: We repeatedly checked Table 4 and 5. Corresponding revisions has been done in revised manuscript.

17.Line 165: BLAST not BLSAT

>>>Response: Thank you for your suggestion, 'BLSAT' had been modified in revised manuscript.

18.Line 175 (and others): keep a space between numbers and units, consistently.

>>>Response: we re-checked all numbers and units throughout the manuscript. All needed spaces have been added between numbers and units.

19.Line195: Change start of sentence (e.g. 'After the above described steps...')

>>> Response: Thank you for your suggestion, corresponding revision has been done in Line 970|Page 8 of the revised manuscript.

20.Line 196: should read: 'than the predicted gene markers..'

>>>Response: Thank you for your suggestion, corresponding revision has been done in Line 971 | Page 8 of revised version.

21.Line203: space between 'by' and 'sequence'

>>>Response: Thank you for your suggestion, a space had been added between by and sequence

22.Line211: after ref 38, just one dot

>>>Response: Thank you for your suggestion, the ref 38 and dot has been deleted in revised version.

23.Line 219: remove space between 'mapping' and ','

>>>Response: Thank you for your suggestion, the space has been deleted between 'mapping' and ','.

24.References: need a lot of editing to uniform

>>>Response: All references of the manuscript have been reviewed and edited based on the author guideline of "Gigascience" in the revised manuscript.

25.Tables: Headers are unclear and many abbreviations within tables are not explained

>>>Response: Thank you for your suggestion, revisions have been done for the table headers. Meanwhile, the abbreviations have been explained and replaced with corresponding full name.

26.What is the difference between Table 4 and Table 7? Both show BUSCO assessments of palm species. Clarify both in tables and in the text.

>>>Response: Thank you for your suggestion, Table 7 has been changed into Table 6 in the revised version. Table 4 referred to the comparative analysis of the assembled genome sequences for four palm species using BUSCO software, while Table 6 referred to the comparative analysis of the predicted gene from the four palm species using BUSCO software. Revisions have been done to make Table 4 and Table 6 legends more clearly in "Table" part of revised version.

27.Figure legend: Figure 1 does not contain any morphological characteristics; they are photographs of coconut plants.

>>>Response: Figure 1 had been substantially revised in the revised version.

Reviewer 2

1.My only major concern about the manuscript is that the written style is not ready for publication. There are many type and grammatical mistakes all over the main text, figure captions and table legends. The manuscript needs some extensive copy editing to be published.

>>>Response: Thank you for your suggestion, the manuscript has been reviewed and edited throughout the manuscript by the native experts (Annaliese Mason, Baudouin Luc and Amjad Iqbal).

Reviewer 3

1. Homologous gene families using a larger set of genomes would allow a gain/-loss analysis (check the *Zostera* (seagrass) genome paper Figure 1a for a recent example), some venn diagrams based on this showing how many gene are shared with close relative (e.g. *Elaeis*), other monocots (e.g. rice) and dicots (e.g. *Arabidopsis*) could also be generated based on this (e.g. orchid genome paper figure 1a). Asynteny/collinearity analysis is usually included, often combined with a Ks analysis (see the orchid genome paper Figure 2, *Zostera* genome paper Figure 2).

>>>Response: Thank you for your suggestion, we added venn diagrams between different species and analyzed the divergence time between different species into Line 990 | Page 8 - Line 1270 | Page 10 of the revised version. Meanwhile, we identified and characterized antiporter and ion channel gene family in Line 1271 | Page 10 – Line 1578 | Page 11 of revised manuscript.

2. No case study is included, I feel there should be at least one (though as the paper is submitted as a data note the journal might not require one). The authors are the first ones to have a glimpse at the genome of this species. I would make sense to check a few relevant gene families (coconut are clearly very different from seeds of other monocots, so seed related gene families would be likely candidates for a more in depth study)

>>>Response: Thank you for your suggestion. It is known that coconut palm can disseminate through ocean currents: floating nuts sprout and grow naturally upon washing up on beaches. The ability to adapt to a high salt environment is closely related to this dissemination feature and to these natural growth conditions. In the revised manuscript, we identified antiporter and ion channel genes in the genome of *Cocos nucifera*, some of which had been validated to be associated with salt stress in *Arabidopsis*. In the gene expansions analysis, some gene families showed significant expansion in compared to *Arabidopsis*, including Na<sup>+</sup>/H<sup>+</sup> antiporter family, Carnitine/acylcarnitine translocase family, Potassium-dependent sodium antiporter, and potassium channel. The expansion of Na<sup>+</sup>/H<sup>+</sup> antiporter family and Potassium-dependent sodium antiporter may be associated with coconut salt tolerance. The expansion of carnitine/acylcarnitine translocase family may be associated with the accumulation of fatty acid in coconut pulp. At last, the expansion of potassium channel may be associated with the accumulation of potassium ion in coconut water. Corresponding revision had been added into Line Line 1271 | Page 10 – Line 1578 | Page 11 of revised manuscript.

3. For non-bioinformaticians a supplemental website which offers a BLAST interface would certainly be welcome.

>>>Response: we have uploaded coconut genome raw data into Sequence Read Archive (SRA) of the National Center for Biotechnology Information. The assembled and annotated data were uploaded into GigaDB database. Meanwhile, the assembled and annotated data have been uploaded into pirate website for blast analysis and genome browse. However, currently, this website is not available for all people. The website will be available after further website improvement and paper publication

4. Line 128 -129: The N50 by itself is not a direct measure for the quality of the assembly. Avoid over-interpretation.

>>>Response: Thank you for your suggestion, the sentence has been replaced by other sentence: "The comparative results of the BUSCO estimation in coconut and in the four other palm genome sequences indicates that the smallest fraction of missing genes as predicted by BUSCO was found in the coconut genome assembly", in Line 724 – Line 726 | Page 6 of revised version.

5. Line 54 and abstract: (DVP01, 4166) -> the number of genes for the date palm

genome is incorrect. In table 2 the authors report 41, 660 !

>>>Response: Thank you for your suggestion; we have re-checked the annotated gene number for datepalm based on the document reported by Al-Mssallem et al., 2013 and 41 660 genes were annotated. The corresponding revisions have been made in the Abstract part of revised manuscript.

6.Line 60: facilitating future: missing space Line 78-79

>>>Response: Thank you for your suggestion, a space has been added into between facilitating and future.

7.Line 78-79: For high tolerant to high salt density: revise grammar

>>>Response: Thank you for your suggestion, revisions have been done in Line 240–Line 242 | Page 3 of revised manuscript

8.Line 80: ...present Hainan Tall... -> ...present the Hainan Tall

>>>Response: Thank you for your suggestion, the sentence has been rewrite and the usage of the phrase “Hainan Tall” has been carefully checked throughout the revised manuscript.

9.Line 82: ...about genome

>>>Response: Revisions have been done in revised manuscript.

10.Line 88 (and other places): ...pair end... -> ...paired end...

>>>Response: Thank you for your suggestion, all ‘pair end’ has been modified into ‘paired end’ throughout the revised manuscript.

11.Line 96: ...removed by using ... -> ...removed using...

>>>Response: Thank you for your suggestion, corresponding revision had been done in revised manuscript

12.Line 116: ...SOAPdenovo2 map... -> ...SOAPdenovo2 maps...

>>>Response: Thank you for your suggestion, corresponding revision had been done in Line 572 | Page 5 of revised manuscript.

13.Line 132: ...reported in previous Fan’s research ....: incorrect grammar should be revised (as previously reported by Fan et al...).

>>>Response: Thank you for your suggestion, corresponding revision had been done in Line 598 | Page 5 in revised manuscript

14.Line 181: Previous Fan’s research: revise grammar

>>>Response: ‘Previous Fan’s research’ had been modified into ‘as previously reported by Fan et al.’ in Line 874 | Page 7 of revised manuscript

15.Line 192: ... a diagrammic pipeline is showed...: revise grammar

>>>Response: Thank you for your suggestion, corresponding revisions had been done in Line 968|Page 8 revised version.

16.Line 199: ...completely... -> complete

>>>Response: Thank you for your suggestion, corresponding revision has been done in Line 973 | Page 8 of revised manuscript.

	<p>17.Line 203 -204: In sequence similarity step: revise</p> <p>&gt;&gt;&gt;Response: 'In sequence similarity step' has been modified into 'Firstly' in Line 979 Page 8 of the revised manuscript.</p> <p>18.Line 232-233: Font is suddenly somewhat bigger</p> <p>&gt;&gt;&gt;Response: Thank you for your suggestion, corresponding revision have been done in "Funding" part of revised manuscript.</p>
<b>Additional Information:</b>	
<b>Question</b>	<b>Response</b>
Are you submitting this manuscript to a special series or article collection?	No
<p><b>Experimental design and statistics</b></p> <p>Full details of the experimental design and statistical methods used should be given in the Methods section, as detailed in our <a href="#">Minimum Standards Reporting Checklist</a>. 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><b>Resources</b></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 <a href="#">Research Resource Identifiers</a> (RRIDs) for antibodies, model organisms and tools, where possible.</p> <p>Have you included the information requested as detailed in our <a href="#">Minimum Standards Reporting Checklist</a>?</p>	Yes
<p><b>Availability of data and materials</b></p> <p>All datasets and code on which the conclusions of the paper rely must be either included in your submission or deposited in <a href="#">publicly available repositories</a> (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.</p>	Yes



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

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

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

2 Yong Xiao<sup>1\*</sup>, Pengwei Xu<sup>3\*</sup>, Haikuo Fan<sup>1\*</sup>, Luc Baudouin<sup>4\*</sup>, Wei Xia<sup>1\*</sup>, Stéphanie Bocs<sup>4\*</sup>, Junyang  
3 Xu<sup>3</sup>, Qiong Li<sup>2</sup>, Anping Guo<sup>2</sup>, Lixia Zhou<sup>1</sup>, Jing Li<sup>1</sup>, Yi Wu<sup>1</sup>, Zilong Ma<sup>2</sup>, Alix Armero<sup>4</sup>, Auguste  
4 Emmanuel Issali<sup>5</sup>, Na Liu<sup>3&</sup>, Ming Peng<sup>2&</sup>, Yaodong Yang<sup>1&</sup>

5 <sup>1</sup>Hainan Key Laboratory of Tropical Oil Crops Biology/Coconut Research Institute, Chinese  
6 Academy of Tropical Agricultural Sciences, Wenchang, Hainan 571339, P.R.China

7 <sup>2</sup>Institute of Tropical Bioscience and Biotechnology, Chinese Academy of Tropical Agricultural  
8 Science, Haikou, Hainan 571101, P. R. China

9 <sup>3</sup>BGI-Shenzhen, Shenzhen 518083, China

10 <sup>4</sup>French Agriculture Research Centre for International for International Development (CIRAD), UMR  
11 AGAP, F-34398, Montpellier France

12 <sup>5</sup>Station Cocotier Marc Delorme, Centre National De RechercheAgronomique (CNRA)\_07 B.P. 13,  
13 Port Bouet,Côte d'Ivoire

15 \*The authors have equal contribution to the manuscript

16 &Corresponding author

17 **Yong Xiao:** [xiaoyong1980@catas.cn](mailto:xiaoyong1980@catas.cn)

18 **Wei Pengwei Xu:** [xupengwei@genomics.cn](mailto:xupengwei@genomics.cn)

19 **Haikuo Fan:** [vanheco@163.com](mailto:vanheco@163.com)

20 **Luc Baudouin:** [luc.baudouin@cirad.fr](mailto:luc.baudouin@cirad.fr)

21 **Wei Xia:** [saizjxiawei@hainu.edu.cn](mailto:saizjxiawei@hainu.edu.cn)

22 **Stéphanie Bocs:** [stephanie.sidibe-bocs@cirad.fr](mailto:stephanie.sidibe-bocs@cirad.fr)

23 **Junyang Xu:** [xujy@genomics.cn](mailto:xujy@genomics.cn)

24 **Qiong Li:** [liqiong4416@126.com](mailto:liqiong4416@126.com)

25 **Anping Guo:** [gap211@126.com](mailto:gap211@126.com)

26 **Lixia Zhou:** [glzz\\_2009@163.com](mailto:glzz_2009@163.com)

27 **Jing Li:** [lijing002x@catas.cn](mailto:lijing002x@catas.cn)

28 **Yi Wu:** [wuyi-scuta@163.com](mailto:wuyi-scuta@163.com)

29 **Zilong Ma:** [mzl900@163.com](mailto:mzl900@163.com)

30

Formatted: Numbering: Continuous

Field Code Changed

Field Code Changed

Formatted: Font: Times New Roman, 9 pt

Formatted: Font: Times New Roman

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 **Armero** ~~Alix~~ **Armero**: [alix.armero\\_villanueva@cirad.fr](mailto:alix.armero_villanueva@cirad.fr)
- 32 **Auguste Emmanuel Issali**: [issaliemma@yahoo.com](mailto:issaliemma@yahoo.com)
- 33 **Na Liu**: [naliu@genomics.cn](mailto:naliu@genomics.cn)
- 34 **Ming Peng**: [mmpeng\\_2000@yahoo.com](mailto:mmpeng_2000@yahoo.com)
- 35 **Yaodong Yang**: [yvjang@catas.cn](mailto:yvjang@catas.cn)

37 **Background**

38 Coconut palm (*Cocos nucifera*, 2n = 32), a member of genus *Cocos* and ~~of the~~ family Arecaceae  
39 (Palmaceae), is an important tropical fruit and oil crop. Currently, coconut palm is cultivated in 93  
40 countries, including Central and South America, East and West Africa, Southeast Asia and the Pacific  
41 island, with a total growth area of more than ~~1212~~ million hectares ([www.fao.org/faostat/en/](http://www.fao.org/faostat/en/)).  
42 Coconut palm is generally classified into two main categories: “Tall” (flowering 8-10 years after  
43 planting) and “Dwarf” (flowering 4-6 years after planting), based on ~~the~~ morphological characteristics  
44 and breeding habits. This ~~palmae~~ ~~Palmae~~ species ~~needs~~ ~~has a~~ long growth ~~time~~ ~~period~~ before ~~entre~~  
45 ~~into~~ reproductive years which hinders ~~the~~ ~~progress~~ ~~for~~ conventional breeding ~~progress~~. In spite of  
46 initial successes, ~~the~~ ~~improvements~~ ~~made by~~ conventional ~~improvement~~ ~~breeding~~ ~~have been~~ ~~is~~ very  
47 ~~slow~~. In the present study, we obtained *de novo* sequences of *Cocos nucifera* genome: ~~which will a~~  
48 ~~major~~ ~~with its~~ ~~provide~~ ~~enormous~~ ~~large~~ ~~amount of~~ genomic information ~~resource~~ which could be used  
49 ~~for~~ ~~to~~ ~~facilitate~~ ~~the~~ ~~further~~ ~~molecular~~ ~~assisted~~ ~~breeding~~ ~~and~~ ~~accelerate~~ ~~the~~ ~~breeding~~ ~~process~~ ~~of~~ ~~in~~ *Cocos*  
50 *nucifera* ~~and~~ ~~accelerating~~ ~~the~~ ~~breeding~~ ~~process~~ ~~in~~ ~~this~~ ~~important~~ ~~crop~~. genetic improvement is very  
51 ~~slow~~. ~~In~~ ~~The~~ ~~present~~ ~~study~~, we obtained ~~the~~ ~~was~~ ~~performed~~ ~~to~~ ~~de~~ ~~novo~~ ~~sequence~~ ~~the~~ ~~genome~~ ~~of~~ *Cocos*  
52 *nucifera* ~~genome~~, which will provide a large amount of genomic information for molecular-assisted  
53 ~~breeding~~ ~~and~~ ~~accelerate~~ ~~the~~ ~~breeding~~ ~~process~~ ~~of~~ *Cocos nucifera*.

54 **Findings**

55 A total of ~~419.67~~ ~~419.67~~ gigabases (Gb) of raw ~~reads~~ ~~reads~~ ~~was~~ ~~were~~ generated by the Illumina HiSeq  
56 2000 platform ~~using~~ ~~different~~ ~~a~~ ~~series~~ ~~of~~ ~~combinations~~ ~~of~~ ~~paired~~ ~~end~~ ~~and~~ ~~mate~~ ~~pair~~ ~~libraries~~ ~~using~~  
57 ~~different~~ ~~combinations~~ ~~of~~ ~~paired~~ ~~end~~ ~~and~~ ~~mate~~ ~~pair~~ ~~libraries~~, ~~comprising~~ ~~which~~ ~~covering~~  
58 ~~approximately~~ ~~173.32~~ ~~depth~~ ~~of~~ ~~the~~ ~~173.32~~ ~~depth~~ ~~of~~ ~~the~~ ~~predicted~~ ~~estimated~~ *Cocos nucifera* ~~genome~~  
59 ~~length~~ *Cocos nucifera* ~~genome~~ (2.42Gb, variety “~~the~~ Hainan Tall”) to an estimated 173.32× read depth.  
60 A total scaffold length of 2.20 Gb was generated ~~with~~ ~~a~~ ~~scaffold~~ ~~N50~~ ~~of~~ ~~418~~ ~~Kb~~, ~~which~~ ~~and~~

Formatted: Font: No underline, Font color: Auto, (Intl) SimSun

Formatted: Font: Times New Roman, 9 pt

Formatted: Font: Times New Roman

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

representings 90.91% of the ~~estimated genome (2.42Gb), while the BUSCO evaluation demonstrated the obtained scaffold sequences reached 90.8% completeness of coconut genome reached (90.8%)~~. The coconut genome was predicted to harbor 28,039 protein-coding genes, which is less than in *Phoenix dactylifera* (PDK30 variety; 28,889), *Phoenix dactylifera* (DPV01 variety; 41,660) and *Elaeis guineensis* (34,802). ~~BUSCO evaluation demonstrated the obtained scaffold sequences covered 90.8% of the coconut genome, and that The completeness level for the annotation completeness genome annotation was also evaluated estimated by BUSCO which showed a level of that was, reached to 74.1% completeness~~. Genome annotation results revealed that 72.75% of the coconut genome ~~was consisted~~ of transposable elements, ~~among inof~~Of which ~~the class of~~ long-terminal repeat ~~retrotransposons~~ elements (LTRs) ~~make upaccounted~~ for the largest proportion (92.23%). ~~Comparative analysis of the antiporter gene family and ion channel gene families between C. nucifera and Arabidopsis thaliana indicated thatasuggested~~ significant gene expansion may ~~happenedoccurred in the former genomecoconut, involving in Na+/H+ antiporter, Carnitine/acylcarnitine translocase, Potassium-dependent sodium-calcium exchanger, and potassium channel genes.~~

### Conclusions

Despite its agronomic importance, ~~Cocos C. nucifera~~ *Cocos nucifera* is still under-studied. ~~In the currentthis report, we made an attempt to construct a draft the genome of Cocos nucifera, whichand provided an enormousa large amount of genomic information that will facilitate future functional genomics and molecular assisted breeding in Cocos nuciferathis crop species.~~ ~~In the current report, we made an attempt to draft the We report a genome draft of Cocosnucifera, which.Thisstudy providesa large amount of genomic information that will, facilitateingfuture functional genomics and molecular assisted breeding inCocosnucifera.~~

**Keywords:** Coconut, ~~Ppalm~~ palm, genome, Assembly, Annotation

### Data description

#### Background

Coconut ~~palm (Cocos nucifera, 2n = 32), the only species of fromin genus Cocos of andand belongs toin the family Arecaceae, is a tropical oil crop and is widely cultivated in tropical regions due to its extensive application in agriculture and industry. The Coconut palm tropical species was is~~ thought to

Formatted: No underline, Font color: Auto

Formatted: Font: Not Italic, No underline, Font color: Auto

Formatted: Font: Not Italic, No underline, Font color: Auto

Formatted: Font: Times New Roman, 9 pt

Formatted: Font: Times New Roman

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

91 be originated from the ~~s~~Southwest and western-Western pacific-Pacific region (including ~~the~~ Malay  
92 Peninsula and ~~arehipelago~~Archipelago, New Guinea, and the Bismarck Archipelago)~~and the~~  
93 southwest Pacific. ~~At P~~Presently, ~~the-this~~ tropical tree crop ~~had-hasis been~~ distributed across 93  
94 tropical countries [1], including Central and South American, East and West African, Southeast Asia  
95 and the ~~pacific-Pacific islands~~Islands, and ~~is grown~~accounts for over 12 million hectares of land  
96 (www.fao.org/faostat/en/).

97 In China, ~~the~~ coconut palm grows in the ~~subtropical regions~~ - Hainan and Yunnan provinces - as  
98 an economic~~al~~ and ornamental plant. ~~In the province of Hainan, e~~Coconut palm is cultivated ~~over~~  
99 ~~forover an area of~~ approximately 43,000 hectares ~~in Hainan, withand, out of which the “Hainan Tall”~~  
100 ~~(HAT) variety covereding~~ approximately 36,000 hectares ~~is made upcovered by the coconut~~  
101 ~~variety, the “theHainan Tall” (HAT) [2].~~ The Hainan TallHAT coconut are ~~needs eight to ten years to~~  
102 ~~entrerites~~ reproductive stage ~~andslow to mature (flowering 8-10 years after planting); has can grow to~~  
103 a height of ~~about~~ 20-30 meters, ~~and have with a~~ medium to large ~~sized~~ nut-size. ~~The Hainan Tall~~  
104 ~~coconut~~ Though ~~this~~ HAT cultivar ~~of coconut is~~ highly tolerant to salt and drought stress, ~~whilebut~~  
105 ~~yet sensitive to~~ temperatures below 10 °C. ~~It is known that under natural conditions, e~~Coconut palm  
106 ~~can can be~~ disseminated through ocean currents: ~~floating on the sea and the nuts that sprouts and~~  
107 ~~grows naturally upon washing up onwhen reach the beach in natural conditiones.~~ The ability ~~ofto~~  
108 ~~adapting to a high salt environment is closely related withto~~ this dissemination feature and to these  
109 ~~natural growth environment~~ Hence, ~~this tropical species gradually adapted to high salt environment~~  
110 ~~during a long evolutionconditionary process.~~ The morphological characteristics of the Hainan  
111 TallHAT cultivar ~~are given~~ shownedn -in Figure 1. Here, ~~Besides, w~~We also present the genome  
112 ~~sequence of Hainan TallHATthe Hainan Tall coconut and thean analysis forof the antiporter and ion~~  
113 ~~channel gene familyfamilies, relevant to salinity tolerance, which will forms the basis for future~~  
114 ~~research investigating the coconuts tolerance to salt stress. Moreover, Since theAs~~ draft genome  
115 ~~sequences of its coconut relative species,s (e.g. such as Elaeis guineensis[3] [3] and Phoenix~~  
116 ~~dactylifera [4, 5], [4, 5]) have previously beenwere also reported-, we also performed TheAa~~  
117 ~~comparative analysis has beenwas performed between coconut and itsthese relative species for the~~  
118 ~~characters of the genome assembly and annotation characteristics results of coconut and its relative~~  
119 ~~species in the study.~~

Formatted: Font: Times New Roman, 9 pt  
Formatted: Font: Times New Roman

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

121 **Data description**

122 **Sample collection and sequence-sequencing strategy**

123 The G genomic DNA was extracted from the spear leaf of the variety a~~the the~~ Hainan Tall” coconut  
124 (*Cocos nucifera* L. Taxonomy ID: 13894; ~~19033’3”~~ 19°33’3”N, 110°047’25”E) individual selected  
125 from the coconut garden of ~~the the~~ Coconut Research Institute (Wenchang, Hainan province, China)  
126 by using the CTAB extraction method [6]. Subsequently, four ~~paired~~ paired-end (PE) libraries with  
127 insert sizes ~~as of~~ 170 bp, 500 bp, 450 bp and 800 bp and five ~~Mate~~ mate-pair (MP) libraries with insert  
128 sizes ~~as of~~ 2 Kb, 5 Kb, 10\_Kb, 20\_Kb and 40 Kb were constructed using the standard procedure  
129 provided by Illumina (San Diego, USA). After library preparation and quality control of the DNA  
130 samples, template DNA fragments were hybridized to the surface of the flow cells on an Illumina  
131 HiSeq2000 sequencer ~~and~~, amplified to form clusters, and then sequenced by following the standard  
132 Illumina manual. Finally, we generated 714.67 Gb of raw reads from all constructed libraries. The  
133 raw ~~outputs for sequeneed~~ outputs for each sequenced library are summarized in Table 1. Before  
134 assembly, ~~We filtered the raw reads were pretreated with~~ using the following stringent filtering  
135 processes ~~throughvia the~~ SOAPfilter (v2.2) [7] software: (1) ~~removed~~ Filtered reads with 25%  
136 low-quality bases (quality scores ≤ 7); (2) Rremoved reads with N bases more than 1%; (3)  
137 Ddiscarded reads with adapter contamination and/or PCR duplicates.; (4) ~~removed~~ Filtered reads with  
138 undersized insert sizes. Finally, 419.08 Gb (estimated 173.17× read depth) of high-quality sequences  
139 were obtained for genome assembly.

140 ~~reads with low quality (base quality less than 7 with percent higher than 25% or N percent higher than~~  
141 ~~1%), small insert size, PCR duplication or adapter contamination were removed using SOAPfilter, a~~  
142 ~~software application in the SOAPdenovo package [7]. After filtering, 419.08 Gb (173.17× depth)~~  
143 ~~high-quality sequences were obtained for genome assembly.~~

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

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

149 
$$G = N \times (L - K + 1) / K\_depth$$

Formatted: No underline, Font color: Auto, Highlight

Formatted: Font: 10 pt

Formatted: Font: Times New Roman, 9 pt

Formatted: Font: Times New Roman

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

150  $G = N \times (L - K + 1) / K\_depth$

151  
152 where N represents the total of number of reads, L represents the read length, K represents the k-mer  
153 value used in the analysis and K\_depth refers to the main peak in the k-mer distribution curve. In our  
154 calculations, N was 2,049,520,223, L was 100 and K\_depth was 71 for K=17. As a result,  
155 therefore, *Cocos nucifera* genome was estimated to be 2.42 gigabases (Gb). K-mer\_size distribution  
156 analysis (Figure 2) indicated that *Cocos nucifera* was a diploid species with low heterozygous  
157 heterozygosity and a high proportion of repetitive sequences.

158 We then assembled the *Cocos nucifera* genome by using the software SOAPdenovo2 in three  
159 steps: contig construction, scaffold construction and gap filling. In the contig construction step: the  
160 SOAPdenovo2 was run with the parameters 'pregraph -K 63 -R -d 1' was employed to construct de  
161 Bruijn graphs from paired-end libraries with an insert size ranging from 170 to 800 bp. Then  
162 the k-mers from the de Bruijn graphs were then used to form contiguous sequences (contigs) with  
163 the parameters 'contig -R' by clipping tips, merging bubbles and removing the low coverage links.  
164 In the scaffold construction step: the orders of the contigs were determined by using paired-end and  
165 mate-pair information with parameters 'map -k 43' and 'scaff -F -u'. In more detail, more detail,  
166 SOAPdenovo2 maps the reads from paired-end and mate pair libraries to contigs based  
167 on a hash table (keys are unique k-mers on contigs; values are positions). In this such cases, two  
168 contigs are considered to be linked if the bridging of the contigs are supported by five paired-end read  
169 pairs or three mate-pair read pairs. In the gap filling step: the bridging of the contigs are supported  
170 by five paired-end read pairs or three mate-pair read pairs. In the gap filling step: gaps within  
171 scaffolds were filled by utilizing KGF [7] (V1.06) and GapCloser software (v1.12-r6) [7] with  
172 paired-end libraries with (having an insert size from 170 to 800 bp in cases, where one end could be  
173 mapped to one contig and the other end extended into a gap). To achieve optimal the assembled  
174 sequence result, Rabbit (a Poisson-based Kk-mer model software, see the URL in the "availability of  
175 supporting data" section path: availability of supporting data software, path: availability of supporting  
176 data) was used to determine repeat sequences, segmental duplications or divergent haplotypes on the  
177 assembly. After removal of the redundant sequences, a final total scaffold length of 2.20 Gb for  
178 the scaffolds was obtained and used for continued further generated analysis, which accountings

Formatted: Font: Times New Roman, 9 pt  
Formatted: Font: Times New Roman

1  
2  
3  
4  
5  
6  
7 179 ~~forecomprising~~ 90.91% of the predicted genome size (Table 2), ~~which was and~~ larger than the African  
8  
9 180 ~~oil palm and date other palm genomes (Table 2) species in palmaceae~~. Meanwhile, ~~the N50 of the the~~  
10  
11 181 ~~obtained contigs N50 was 72.64 Kb and the scaffold N50 was 418.06 Kb for the scaffolds, which~~  
12  
13 182 ~~have excluded~~ ~~while the length of scaffolds less than 100 bp were excluded~~. The ~~C~~comparison of  
14  
15 183 ~~N50 values for the s-assembled y-N50s of~~ coconut genome and for with four previously published  
16  
17 184 palm genomes *Elaeis guineensis* [3], *Elaeis oleifera* [3], *Phoenix dactylifera* (PDK30) [4] and  
18  
19 185 *Phoenix dactylifera* (DPV01) [5] ~~were listed in~~ were listed in Table 2.

Formatted: No underline, Font color: Auto, Not Highlight

Formatted: No underline, Font color: Auto, Not Highlight

## 186 Genome evaluation

187 The 57,304 unigenes (transcript obtained from three different tissues, spear leaves, young leaves and  
188 fruit flesh) ~~as previously reported by Fan et al. reported by Fan et al.~~ [8] were aligned to the assembled  
189 genome of *Cocos nucifera* using BLAT [9] with default parameters. The alignment results ~~predicted~~  
190 ~~indicated~~ that the assembled genome of *Cocos nucifera* covered 96.78% of ~~all the~~ expressed unigenes,  
191 suggesting a high level of coverage. ~~has been reached for the assembled genome~~ (Table 3).

192 We also evaluated the ~~level of genome completeness~~ ~~of~~ the assembled sequences ~~by~~ using  
193 BUSCOv2.0 [10], which ~~quantitatively assesses~~ genome completeness ~~by~~ using  
194 evolutionarily-informed expectations of gene content from near-universal single-copy orthologs  
195 selected from OrthoDBv9 (<http://busco.ezlab.org/>, plant set). BUSCO analysis showed that ~~there are~~  
196 ~~separate~~ 90.8% and 3.4% of the 1,440 expected plant genes were identified as complete and  
197 fragmented ~~genes respectively, respectively~~, while 5.8% ~~of genes~~ were considered ~~as to be~~ missing ~~in~~  
198 ~~from the assembled coconut genome sequence~~. The BUSCO results ~~showed that our assembly was~~  
199 ~~more complete than assembled data reported from three palm species. The Comparison~~  
200 ~~C comparative results of the BUSCO result estimation with in coconut and in the the other four~~  
201 ~~other palm genome sequences~~ ~~indicated~~ ~~that the smallest fraction of missing genes as predicted by~~  
202 ~~BUSCO genes happened~~ was found in the coconut genome ~~assembl~~ ~~assembly~~ (Table 4). ~~the BUSCO~~  
203 ~~results with the other four palm genomes indicated the smallest missing of smallest BUSCO genes in~~  
204 ~~coconut genome (Table 4)~~.

Formatted: No underline, Font color: Auto, Not Highlight

Formatted: No underline, Font color: Auto, Not Highlight

Formatted: No underline, Font color: Auto, Not Highlight

Formatted: No underline, Font color: Auto, Not Highlight

Formatted: No underline, Font color: Auto, Not Highlight

Formatted: No underline, Font color: Auto, Not Highlight

Formatted: No underline, Font color: Auto, Not Highlight

Formatted: No underline, Font color: Auto, Not Highlight

Formatted: No underline, Font color: Auto, Not Highlight

Formatted: No underline, Font color: Auto, Not Highlight

Formatted: No underline, Font color: Auto, Not Highlight

Formatted: No underline, Font color: Auto, Not Highlight

Formatted: Font: Times New Roman, 9 pt

Formatted: Font: Times New Roman

## 205 Repeat annotation

206 We combined ~~a~~-homology ~~- based annotation~~ and *de novo* method to identify transposable elements  
207 (TEs) and the tandem repeats in the *Cocos nucifera* genome. In homology ~~- based annotation~~ step:  
208 TEs ~~at DNA and protein levels~~ were identified by searching against ~~the~~ Repbase library\_ (version  
2

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



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

209 20.04) [11] with RepeatMasker (v4.0.5) [12] and RepeatProteinMasker (v4.0.5) [12]. In *the de novo*  
210 step, *de novo* libraries were constructed based on the genome sequences using the *de novo* prediction  
211 program RepeatModeler (~~path: See the URL in the “availability of supporting data” section~~ [availability of](#)  
212 [supporting data” section](#)) and LTR\_FINDER [13] by removing ~~contamination-contaminant~~ and  
213 multi-copy genes. ~~Subsequently, Then the~~ novel transposable elements were identified and classified  
214 using RepeatMasker. ~~The~~ tandem repeat sequences were identified by TRF (Tandem Repeat Finder)  
215 software [14] with the following parameters ‘Match=<sub>2</sub>, Mismatch=<sub>7</sub>, Delta=<sub>7</sub>, PM=<sub>80</sub>, PI=<sub>10</sub>,  
216 Minscore=<sub>50</sub> and MaxPeriod=<sub>2000</sub>’. The total length of the tandem repeat sequences predicted by  
217 the software ~~is was~~ 151,229,585 bp, comprising 6.86% of the coconut genome. Finally, ~~a total of~~ 1.6  
218 Gb of non-redundant repetitive elements were identified, accounting for 74.48% of the coconut  
219 genome, ~~while~~ Transposable elements took up 72.75% ~~for of the total 1.6Gb of repetitive elements~~  
220 ~~and with the~~ The most predominant transposons were long-terminal repeat [retrotransposon](#) (LTR)  
221 ~~class, which~~ [sing](#) for 92.23% of all TEs and 67.1% of the coconut genome (Table 5).

## 222 Gene prediction

223 ~~We combined homology, de novo and transcript alignment to predict genes in Cocos nucifera genome.~~  
224 We combined ~~three strategies to predict genes in Cocos nucifera genome:~~ homology -based, *de novo*  
225 and transcript alignment ~~to predict genes in Cocos nucifera genome.~~ For homology prediction: For  
226 ~~homology prediction- based annotation:~~ the protein sequences ~~the protein sequences of Arabidopsis~~  
227 ~~thaliana [15], Oryza sativa [16], Sorghum bicolor [17] and Zea mays [18] of Arabidopsis thaliana [15],~~  
228 ~~Oryza sativa [16], Sorghum bicolor [17], Zea mays [18], Elaeis guineensis, and Phoenix dactylifera~~  
229 ~~(DPV01) and Elaeis guineensis and Phoenix dactylifera (DPV01)~~ were downloaded from each  
230 ~~corresponding sources (see “Availability of data sources”) from each corresponding sources (See~~  
231 ~~“Availability of data sources”). The longest transcript was selected to represent the genes with among~~  
232 ~~differnt alternative splicing variants. We aligned these homologous proteins to t~~ The coconut genome  
233 ~~was blast aligned against these downloaded databases using TBLASTN [19] with parameter ‘-e 1e-5 -F~~  
234 ~~-m 8’; and connected the BLAST hit results were processed to candidate gene loci by solar (v0.9)~~  
235 with parameter ‘-aprot\_2\_genome2 -z’ ~~to determine the candidate gene loci.~~ Next, we extracted the  
236 genomic sequences of candidate gene loci ~~along with up-and-down-stream-1kb~~ flanking sequences,  
237 ~~and applied~~ Genewise 2.2.0 [20] to define the intron - exon ~~boundary boundaries.~~ The genes with  
238 pre-stop codon or frame ~~shifted shifts~~ were excluded ~~for from~~ further analysis.

Formatted: Font: Times New Roman, 9 pt

Formatted: Font: Times New Roman

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

239 For *De-de novo* prediction: ~~We-we~~ randomly selected 1000 ~~full-full~~-length genes (GeneWise  
240 score equal 100, intact structure: start codon, stop codon, perfect intron-exon boundary) from gene  
241 ~~sets-models~~ predicted by ~~homology-homology-based~~ methods to train the model parameters for  
242 AUGUSTUS2.5[21]. Two software programs, AUGUSTU2.5 and GENSCAN 1.0 [22], were used to  
243 do *de novo* prediction on ~~the~~ repeat-masked genome of *Cocos\_nucifera*. Genes with incomplete  
244 structure or protein coding length less than 150bp were filtered out.

245 ~~Subsequently, Then G~~ genes from ~~both~~ homology-based and *de novo* methods were combined to  
246 ~~get-obtain~~ non-redundant gene sets by using GLEAN [23] with the following parameters: minimum  
247 coding sequence length 150 bp and maximum intron length 50 kb. Genes were filtered with the same  
248 thresholds ~~as were used~~ for homology-based annotation.

249 For transcriptome-based prediction: RNA-seq data (SRR606452) ~~as as-previously reported by~~  
250 ~~Fan et al. as prvioiuslypreviously reported by Fan et al.[8] were-was~~ mapped onto the coconut  
251 genome to identify the splice junctions using the software TopHat (v2.1.1) [24]. ~~And thenThe~~  
252 ~~software~~ Cufflinks (v2.2.1) [25] was ~~then~~ used to assemble transcripts with the aligned reads. The  
253 coding potential of these transcripts was identified using ~~a~~ fifth-order Hidden Markov Model, which  
254 was estimated with the same gene sets used in AUGUSTUS training by train\_GlimmerHMM, ~~an~~  
255 application in the GlimmerHMM [26] package. The transcripts with intact open reading frames  
256 (ORFs) were ~~exacted-extracted~~ and the longest ~~ORF-transcript~~ was retrieved ~~as whilea representative~~  
257 ~~of a gene whiles~~ multiple ~~isoformstranscripts fromloeated in-on thea~~ same locus.

258 ~~At lastFinally,~~ we merged the GLEAN and the transcriptome result to form a comprehensive  
259 gene set using an in-house annotation pipeline ~~with the in~~-following steps: firstly, all-to-all  
260 ~~BLASTP analysis of protein sequences wereas~~ performed between GLEAN results and transcript  
261 ~~assemblies with an E-value cutoff of 1e-10. These transcript assemblies were added to the~~  
262 ~~GLEAN result to form (untranslated region) UTRs or alternative spliceing products, depending~~  
263 ~~on whether the coverage and identity of the alignment results reached 0.9 or not. If the transcript~~  
264 ~~assemblies had no blastBLAST hit with the GLEAN results, these transcript assemblies would~~  
265 ~~be~~were added to the final gene set as novel gene. ~~inthe~~following steps: firstly, all-to-all BLASTP  
266 ~~analysis of protein sequences were performed between GLEAN result and transcript assemblies~~  
267 ~~with an E-value cutoff 1e-10. These transcript assemblies were added to the GLEAN result to~~  
268 ~~form (untranslated region) UTRs or alternative splice on whether coverage and identity of the~~

Formatted: Font: Times New Roman, 9 pt  
Formatted: Font: Times New Roman

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

269 alignment results reached 0.9 or not. If the transcript assemblies had no blast hit with the  
270 GLEAN result, these transcript assemblies would be added to the final gene set as novel gene.  
271 The protocol for integrating GLEAN and ~~Transcriptome-transcriptome~~ data is shown in Figure 3.

## 272 Gene evaluation

273 After the above described steps, we obtained a final gene set contained ~~The annotation results~~  
274 ~~showed processes have identified a total of~~ 28,039 protein-coding ~~protein-coding~~ genes ~~were obtained~~  
275 (Table 2), which is less than ~~the predicted-predicted~~ gene numbers of *Phoenix dactylifera*  
276 (PDK30,28,889), *Phoenix dactylifera* (DPV01, 41,660) and *Elaeis guineensis* (34,802). Meanwhile,  
277 ~~the through the~~ BUSCO evaluation ~~showed the a separate of that~~ ~~demonstrated~~ 74.1% and 11.2% of  
278 1,440 expected plant genes were identified as complete and fragmented, ~~and with~~ 14.7% of genes  
279 ~~were~~ considered missing in the gene sets. The BUSCO results showed that our gene prediction was  
280 more complete than ~~that of~~ *Phoenix dactylifera* (PDK30) and *Elaeis guineensis*, ~~but~~ less completely  
281 than ~~that of~~ *Phoenix dactylifera* (DPV01) (Table 6). ~~This maybe due to the higher repetitive elements~~  
282 ~~hinderence of the gene prediction of coconut genome by higher repetitive elements.~~

## 283 Gene Function

284 Gene function annotation was ~~identified done by based on~~ sequence similarity and domains  
285 conservation. ~~In Firstly the step of sequence alignment: we searched aligned~~ the coconut protein  
286 coding genes ~~were blast aligned against against with the~~ KEGG protein ~~databases~~ [27], SwissProt and  
287 TrEMBL [28] using BLASTP at a cut-off E-value threshold of  $10^{-5}$ . ~~Subsequently, the Then we use~~  
288 ~~the best match of from the~~ alignment ~~was used~~ to represent the gene function. We obtained 18,445  
289 KEGG, 18,867 Swissprot and 24,882 Tremble\_annotated genes. ~~In domains conservation~~  
290 ~~step: Secondly, InterProScan\_5.11-51.0 software [29]~~ was employed to identify the motif and domain  
291 ~~based on against~~ the public databases Pfam [30], PRINTS [31], ProDom [32], SMART [33],  
292 PANTHER [34], TIGRFAM [35] and SUPERFAMILY [36]. ~~This The gene function~~  
293 ~~annotation revealed demonstrated demonstrated~~ that 21,087 of the coconut proteins had conserved  
294 motifs ~~and~~ 1,622 Gene Ontology (GO) terms were assigned to 15,705 coconut proteins from the  
295 corresponding InterPro entry [37]. In total, approximately 89.41% of these genes were functionally  
296 annotated using ~~the~~ above methods.

## 297 Gene Family Construction

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

298 Protein sequences of thirteen angiosperms, including *Elaeis guineensis*, *Phoenix dactylifera* (DPV01),  
299 *Sorghum bicolor*, *Prunus persica*, *Solanum tuberosum*, *Glycine max*, *Arabidopsis thaliana*,  
300 *Theobroma cacao*, *Vitis vinifera*, *Musa acuminata*, *Carica papaya*, *Populus trichocarpa* and  
301 *Amborella trichopoda*, were download from each corresponding ftp site Protein sequences of thirteen  
302 angiosperms, including *Elaeis guineensis*, *Phoenix dactylifera* (DPV01), *Sorghum bicolor*, *Prunus*  
303 *persica*, *Solanum tuberosum*, *Glycine max*, *Arabidopsis thaliana*, *Theobroma cacao*, *Vitis vinifera*,  
304 *Musa acuminata*, *Carica papaya*, *Populus trichocarpa*, *Amborella trichopoda*, were download from  
305 each corresponding ftp site (see “Availability of data sources”). For genes with alternative splicing  
306 variants, the longest transcripts were selected to represent the gene. The gene numbers of *Elaeis*  
307 *guineensis* and *Phoenix dactylifera* (DPV01) were greatly different from the research paper published  
308 in 2013 [3, 5] (reference), because genes of these two species were re-predicted using the NCBI  
309 Prokaryotic Genome Annotation Pipeline which seemed to be more reasonable. Similarities between  
310 paired sequences were calculated using BLASTP with an E-value threshold of 1e-5. OrthoMCL [38]  
311 was used to identify gene family based on the similarities of the genes and a Markov Chain Clustering  
312 (MCL) with default parameters For genes with alternative splicing variants, the longest transcript was  
313 selected to represent the gene. The gene numbers of *Elaeis guineensis* and *Phoenix dactylifera*  
314 (DPV01) were greatly different from the research paper published in 2013, because genes of these  
315 two species were re-predicted using NCBI Prokaryotic Genome Annotation Pipeline, which seemed  
316 to be more reasonable. Similarities between pair sequence were calculated using BLASTP with  
317 E-value threshold of 1e-5. OrthoMCL [38] was used to identify gene family based on the similarities  
318 of the genes and a Markov Chain Clustering (MCL) with default parameters. About 79.80% of *Cocos*  
319 *nucifera* genes were assigned into 14,411 families, of which 282 families were only existing in  
320 *Cocos nucifera* (coconut specific families) (Table 7). Figure 4 shows the shared gene families  
321 for orthologous genes. There are 544 orthologous families shared by five monocot species and 7706  
322 orthologous families shared by all monocot and dicot species, suggesting 544 monocot unique  
323 functions shared by five monocot species and 7,706 ancestral functions in the most recent common  
324 ancestor of the angiosperms. of *Cocos nucifera* genes were assigned in 14,411 families, 282 families  
325 were only existing in *Cocos nucifera* (coconut specific families) (Table 7). Figure 4 shows shared gene  
326 families of orthologous genes. There are 544 orthologous families shared by five monocot species and  
327 7706 orthologous families shared by all monocot and dicot species, suggesting 544 monocot unique

Formatted: No underline, Font color: Auto, Highlight

Formatted: Font: Times New Roman, 9 pt

Formatted: Font: Times New Roman

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

328 functions shared by five monocot species and 7,706 ancestral functions in most recent common  
329 ancestor of angiosperms.

### 330 **Phylogenetic analysis**

331 We extracted 247 single copy orthologous genes derived from the gene family analysis step, and  
332 then aligned the protein sequences of each family with MUSCLE (v3.8.31) [39]. Next, the protein  
333 alignments were converted to corresponding coding sequences (CDS) using an in-house Perl script.  
334 These coding sequences of each single copy gene family were concatenated to form one super gene  
335 for each species. The nucleotides at position 2 (phase one site) and 3 (four degenerate site) of codon  
336 were extracted separately to construct the phylogenetic tree by PhyML3.0 [40] withusing a HKY85  
337 substitution model and a gamma distribution across sites. The tree constructed by phase one sites was  
338 consistent with the tree constructed by four degenerate sites.

339 We extracted 247single copy orthologous genes from the gene family step, and then aligned the  
340 protein sequences of each family with MUSCLE (v3.8.31) [39]. Next, the protein alignments were  
341 converted to corresponding coding sequences (CDS) using an in-house Perl script. These coding  
342 sequences of each single copy family were concatenated to form one super gene for each species. The  
343 nucleotides at position 2 (phase one site) and 3 (four degenerate site) of codon were extracted  
344 separately to construct the phylogenetic tree by PhyML3.0 [40] with HKY85 substitution model and a  
345 gamma distribution across sites. The tree constructed by phase one sites was consistent with tree  
346 constructed by four degenerate sites.

### 347 **Divergence time**

348 The Bayesian relaxed molecular clock approach was used to estimate species divergence time using  
349 MCMCTREE in PAML. The Bayesian relaxed molecular clock approach was used to estimate species  
350 divergence time using MCMCTREE in PAML [41], based on the four-degenerate sites based on the  
351 four degenerate sites and the data set used in phylogenetic analysis, with previously published  
352 calibration times [42] (Divergence between *Arabidopsis* between *Arabidopsis thaliana* and *Carica*  
353 *papaya* was 54-90 Mya, divergence between *Arabidopsis thaliana* and *Populus trichocarpa* was  
354 100-120 Mya). The divergence time between coconut and oil palm is about 46.0 (25.4-83.3) million  
355 years ago (Figure 5), which is less than the divergence time between coconut and date palm data set  
356 used in phylogenetic analysis, with previously published calibration times [42] (Divergence

Formatted: Font: Times New Roman, 9 pt

Formatted: Font: Times New Roman

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

357 between *Arabidopsis thaliana* and *Carica papaya* was 54–90 Mya, divergence between *Arabidopsis*  
358 *thaliana* and *Populus trichocarpa* was 100–120 Mya). The divergence time between coconut and oil  
359 palm is about 46.0 (25.4–83.3) million years ago (Figure 5), which is less than the divergence time  
360 between coconut and date palm.

### 361 Identification of antiporter genes in coconut genome

362 Antiporters are a transmembrane proteins involved in the exchange of two substances within and  
363 outside opposite directions through the membrane. In *Arabidopsis*, the functions of *Arabidopsis*  
364 antiporter genes have been well characterized experimentally, and this gene family was  
365 subdivided into thirteen different functional groups. Among them, three functional clusters involved  
366 in Na<sup>+</sup>/H<sup>+</sup> antiporters, some of which were documented to be associated with salt tolerance [43, 44].

Formatted: No underline, Font color: Auto, Not Highlight

367 Antiporter is a transmembrane protein involving in exchange of two substances in opposite  
368 direction through the membrane. In *Arabidopsis*, the functions of *Arabidopsis* antiporter genes have  
369 been well characterized experimentally, and were subdivided into thirteen different functional groups.  
370 Among them, three functional clusters involved in Na<sup>+</sup>/H<sup>+</sup> antiporter, some of which were  
371 documented to be associated with salt tolerance [43, 44].

372 The amino acid sequences of 70 antiporter genes of *Arabidopsis* were downloaded from the  
373 *Arabidopsis* Information Resource (TAIR) website (<http://www.arabidopsis.org>) and used as  
374 queries for BLASTP against the predicated protein databases of the *Cocos nucifera* genome  
375 with a cut-off e-value of 1e-10. A total of 126 antiporter genes were identified in coconut  
376 genome. The amino acid sequences of 70 antiporter genes of *Arabidopsis* downloaded from the  
377 *Arabidopsis* Information Resource (TAIR) website (<http://www.arabidopsis.org>) were used as  
378 queries to BLASTP against the protein database of *Cocos nucifera* at a cut-off e-value of 1e-10. A  
379 total of 126 antiporter genes were identified in coconut genome. With the help of the Using local  
380 Hidden Markov Model-based HMMER (v3.0) searches and the Pfam database, seven antiporter genes  
381 were excluded from further analysis because of the lack of conserved domain. The detailed  
382 information of the 119 antiporter genes were listed in Additional file 1.

383 local Hidden Markov Model-based HMMER (v3.0) searches and Pfam database, seven antiporter  
384 genes were excluded for further analysis because of lack of conserved domain. The detailed  
385 information of the 119 antiporter genes were listed in Supplementary Table 1.

386 In order to elucidate the evolutionary relationship and potential functions of the antiporters

Formatted: Font: Times New Roman, 9 pt

Formatted: Font: Times New Roman

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

387 identified in the study, we applied a combined phylogenetic analysis of *Arabidopsis* and *C. nucifera*  
388 antiporter proteins using the neighbor joining method (Figure 6). Phylogenetic analysis  
389 showed that the 119 antiporter genes from *C. nucifera* can be subdivided into twelve groups.  
390 Meanwhile, and that almost all antiporter genes from *C. nucifera* can be clustered together  
391 with the functional groups found in *Arabidopsis thaliana*.

392 In order to elucidate the evolutionary relationship and potential functions of the EgMYBs  
393 identified in the study, we applied a combined phylogenetic analysis of *Arabidopsis* and *C.*  
394 *nucifera* antiporter proteins using the neighbor joining method (Figure 6). Phylogenetic analysis  
395 showed that the 119 antiporter genes from *C. nucifera* can be subdivided into twelve groups.  
396 Meanwhile, almost all antiporter genes from *C. nucifera* can be clustered into the function groups  
397 found in *Arabidopsis thaliana*.

398 Phylogenetic analysis showed that the number of antiporter genes were equal between  
399 *Arabidopsis thaliana* and *C. nucifera* *Elaeis guineensis* among for almost groups except for G1 (one  
400 of three Na<sup>+</sup>/H<sup>+</sup> antiporter family), G3 (Carnitine/acylcarnitine translocase family) and G12  
401 (Potassium-dependent sodium-calcium exchanger). In the three groups, the genes from *C.*  
402 *nucifera* are far more than these from *Arabidopsis thaliana*, for example, G1 group (one of three  
403 Na<sup>+</sup>/H<sup>+</sup> antiporter families) only contained one *Arabidopsis* antiporter gene and but 14 *C.*  
404 *nucifera* antiporters (1-At/14-Cn). Phylogenetic analysis showed that the number of antiporter genes  
405 were equal between *Arabidopsis thaliana* and *Elaeis guineensis* among all groups except for G1 (one  
406 of three Na<sup>+</sup>/H<sup>+</sup> antiporter family), G3 (Carnitine/acylcarnitine translocase family) and G12  
407 (Potassium-dependent sodium-calcium exchanger). In the three groups, the genes from *C. nucifera* are  
408 far more than these from *Arabidopsis thaliana*, for example, G1 group (one of three Na<sup>+</sup>/H<sup>+</sup>  
409 antiporter family) only contained one *Arabidopsis* antiporter gene, and but 14 *C. nucifera* antiporters  
410 (1/14), whereas G3 (Carnitine/acylcarnitine translocase family) G3 (Carnitine/acylcarnitine  
411 translocase family) contained (1-At/29-Cn), and G13 (PotassiumPotassium-dependent  
412 sodium-calcium exchanger) G3 (PotassiumPotassium-dependent sodium-calcium exchanger)  
413 contained (3-At/11-Cn). These results indicated gene family expansion involving in the three  
414 functional groups. Na<sup>+</sup>/H<sup>+</sup> antiporter family had been reported to be associated with salt stress.  
415 Hence, the expansion of the Na<sup>+</sup>/H<sup>+</sup> antiporter gene family in coconut palm maybe associated with  
416 the high salt tolerance of coconut. Meanwhile, carnitine/acylcarnitine translocase is involved in fatty

Formatted: Font: Times New Roman, 9 pt  
Formatted: Font: Times New Roman

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

417 acid transport cross the mitochondrial membranes. Hence, ~~†~~These gene family expansion of the gene  
418 family maybe associated with accumulation of fatty acid in coconut pulp. Moreover, coconut water  
419 contains a high density of potassium ion, approximately 312 mg potassium ion per 100 g coconut  
420 water [45]. In theis study, the gene number of potassiumpotassium dependent sodium calcium  
421 exchangerswere also detected to be significant expansion. potassium-dependent sodium-calcium  
422 exchangers were also detected to be significantly increased comparinged withto  
423 Arabidopsisexpansion.

424 **Identification of ion channel genes in coconut genome**

425  
426 A total of 67 ion channel genes were identified in the coconut genome (Additional file 2). The amino  
427 acid sequences of 67 *C. nucifera* and 60 *Arabidopsis* ion channel genes were used to  
428 demonstrateanalyze their evolutionary relationship (Figure 7). Almost all ion channel genes from *C.*  
429 *nucifera* can be clustered into the function groups found in *Arabidopsis thaliana*. The number of ion  
430 channel genes was equal between *Arabidopsis thaliana* and *Cocos nucifera* among allin most groups  
431 except for G5 (potassium channel). Many moreThe genes (21) from *C. nucifera* are far more than  
432 these (9) from *Arabidopsis thaliana* (9 genes) were present in group 5 (potassium channels),  
433 whichindicating gene family expansion are involveding in potassium channel. The gene family  
434 expansion maybe associated with accumulation of potassium ions in coconut water.

435 **Conclusion**

436 *Cocos nucifera* (2n = 32) is an important tropical crop, and is also used as an ornamental plant in  
437 the tropics. In the present study, we sequenced and *de novo* assembled the coconut genome. A total  
438 scaffold length of 2.2 Gb was generated, with a-scaffold N50 of 418 Kb. The divergence time of  
439 *Cocos nucifera* and *Elaeis guineensis* is lessmore recent than that of *Cocos nucifera* and *Phoenix*  
440 *dactylifera*, suggesting thea closer relationship ofbetween *C. nucifera* and *E. guineensis* is more  
441 closely. Comparative analysis of antiporter and ion channels between *C. nucifera* and *Arabidopsis*  
442 *thaliana* showed significant gene family expansions maybe involving Na<sup>+</sup>/H<sup>+</sup> antiporters,  
443 carnitine/acylcarnitine translocases, potassium-dependent sodium-calcium exchangers, and potassium  
444 channels. The expansion of these gene families may be associated with adaptation to salt stress,  
445 accumulation of fatty acid in coconut pulp and potassium ions in coconut water.The function of  
446 expanded gene families in species evolution is always tend toalways related towith the environmental

Formatted: Font: Times New Roman, 9 pt  
Formatted: Font: Times New Roman



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

447 ~~adaption species adapt and the identified, the expand gene families of expansion happened in~~  
448 ~~coconut palm may be associated with its' salty adaption and unique taste of coconut water. The~~  
449 ~~divergence time of *Cocos nucifera* and *Elaeis guineensis* is less than *Cocos nucifera* and *Phoenix*~~  
450 ~~*dactylifera*, suggesting the a closer relationship of between *C. nucifera* and *E. guineensis* is more~~  
451 ~~closely. The function of expanded gene families is always related to the environment that a species~~  
452 ~~adapt, therefore, expand gene families of coconut may associate with its' salty adaption and unique~~  
453 ~~taste of coconut water. The data output of the coconut genome will provide a valuable resource and~~  
454 ~~reference information for the development of high density molecular makers, construction of high~~  
455 ~~density linkage maps, detection of QTL (quantitative trait loci), genome-wide association mapping,~~  
456 ~~and molecular breeding. *Comparative analysis of antiporter and ion channel between *C. nucifera* and*~~  
457 ~~*Arabidopsis thaliana* suggested showed significant gene expansion involving in Na<sup>+</sup>/H<sup>+</sup> antiporter,~~  
458 ~~*Carnitine/acylearnitine translocase, Patassium dependent sodium calcium exchanger, and potassium*~~  
459 ~~*channel. The expansion of these gene families may be associated with coconut salt stress,*~~  
460 ~~*accumulation of fatty acid in coconut pulp and potassium ion in coconut water.*~~

461 **Availability of supporting data**

462 Supporting data are available in the GigaDB database, and the raw data were deposited in the  
463 SRA539146 with the project accession code PRJNA374600 for the *Cocos nucifera* genome.  
464 Previously published RNA-seq data used for transcriptome-based prediction is available from the  
465 under accession number SRR606452.

466 **Availability of software**

467 [Rabbit:ftp://ftp.genomics.org.cn/pub/Plutellaxylostella/Rabbit\\_linux-2.6.18-194.blc.tar.gz](ftp://ftp.genomics.org.cn/pub/Plutellaxylostella/Rabbit_linux-2.6.18-194.blc.tar.gz)  
468 [RepeatModeler: http://www.repeatmasker.org/RepeatModeler.html, version 1.0.5](http://www.repeatmasker.org/RepeatModeler.html)  
469 [Solar: https://sourceforge.net/p/treesoft/code/HEAD/tree/branches/lh3/solar/](https://sourceforge.net/p/treesoft/code/HEAD/tree/branches/lh3/solar/)  
470 [HMMER: http://www.ebi.ac.uk/Tools/hmmer](http://www.ebi.ac.uk/Tools/hmmer)

471 **Availability of software**

472 [Rabbit:ftp://ftp.genomics.org.cn/pub/Plutellaxylostella/Rabbit\\_linux\\_2.6.18-194.blc.tar.gz](ftp://ftp.genomics.org.cn/pub/Plutellaxylostella/Rabbit_linux_2.6.18-194.blc.tar.gz)  
473 [RepeatModeler: http://www.repeatmasker.org/RepeatModeler.html, version 1.0.5](http://www.repeatmasker.org/RepeatModeler.html)  
474 [Solar: https://sourceforge.net/p/treesoft/code/HEAD/tree/branches/lh3/solar/](https://sourceforge.net/p/treesoft/code/HEAD/tree/branches/lh3/solar/)  
475 [HMMER: http://www.ebi.ac.uk/Tools/hmmer](http://www.ebi.ac.uk/Tools/hmmer)

Formatted: Font: Italic, No underline, Font color: Auto

Formatted: Font: Italic, No underline, Font color: Auto

Formatted: Line spacing: 1.5 lines

Formatted: Font: Times New Roman, 9 pt

Formatted: Font: Times New Roman

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

476 **Availability of other angiosperms data sources**

477 *Arabidopsis thaliana*, *Oryza sativa*, *Sorghum bicolor*, *Zea mays*, *Sorghum bicolor*, *Solanum*

478 *tuberosum*, *Prunus persica*, *Theobroma cacao*, *Vitis vinifera*, *Musa acuminata*, *Carica papaya*,

479 *Populus trichocarpa*, *Amborella trichopoda*: <https://phytozome.jgi.doe.gov/pz/portal.html>

480 ([phytozomev9.1](#))

481 *Elaeis guineensis*: [ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/442/705/GCF\\_000442705.1\\_EG5/](ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/442/705/GCF_000442705.1_EG5/)

482 *Phoenix dactylifera* (DPV01):

483 [ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/413/155/GCF\\_000413155.1\\_DPV01/](ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/413/155/GCF_000413155.1_DPV01/)

484 *Phoenix dactylifera* (PDK30):

485 <http://qatar-weill.cornell.edu/research/datepalmGenome/download.html>

486 **Availability of other angiosperms data sources**

487 *Arabidopsis thaliana*, *Oryza sativa*, *Sorghum bicolor*, *Zea mays*, *Sorghum bicolor*, *Solanum*

488 *tuberosum*, *Prunus persica*, *Theobroma cacao*, *Vitis vinifera*, *Musa acuminata*, *Carica papaya*,

489 *Populus trichocarpa*, *Amborella trichopoda*: <https://phytozome.jgi.doe.gov/pz/portal.html>

490 ([phytozomev9.1](#))

491 *Elaeis guineensis*: [ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/442/705/GCF\\_000442705.1\\_EG5/](ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/442/705/GCF_000442705.1_EG5/)

492 *Phoenix dactylifera* (DPV01):

493 [ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/413/155/GCF\\_000413155.1\\_DPV01/](ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/413/155/GCF_000413155.1_DPV01/)

494 *Phoenix dactylifera* (PDK30):

495 <http://qatar-weill.cornell.edu/research/datepalmGenome/download.html>

496 **Competing interests**

497 The authors declare that they have no competing interests.

498 **Funding**

499 This study was supported by International Science and Technology Cooperation projects of Hainan

500 Province (No. KJHZ2014-24), Hainan Natural Science Foundation (No. 313058), the major

501 Technology Project of Hainan (No. ZDZX2013023-1), the fundamental Scientific Research Funds for

502 Chinese Academy of Tropical Agriculture Sciences (CATAS-No. 1630032012044, 1630052014002,

503 1630052015050, 1630152017019, and 1630152016006), Central Public-interest Scientific Institution

504 Basal Research Fund for Innovative Research Team Program of CATAS (No. 17CXTD-28).

Formatted: Line spacing: 1.5 lines

Formatted: Left, Line spacing: 1.5 lines

Formatted: Font: Times New Roman, 9 pt

Formatted: Font: Times New Roman

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

505 **Author's contribution**

506 YX, HF, YY, MP, QL, AG designed the study and contribute to the project coordination; XY, PX, WX  
507 wrote the paper; LZ, JL, YW collected the samples and extracted the genomic DNA; YX, BL, BS, JX,  
508 AA, EI, NL conducted the genome analyses.

509 **Acknowledgements**

510 Annaliese S. Mason is gratefully acknowledged for assistance with language editing and manuscript  
511 revisions.

512 **References**

513 1. Batugal P, V Ramanatha Rao and J Oliver, editors. Coconut Genetic Resources. International  
514 Plant Genetic Resources Institute – Regional Office for Asia, the Pacific and Oceania  
515 (IPGRI-APO) Serdang, Selangor DE, Malaysia; 2005.  
516 2. Tang B, Tang M, Chen C, Qiu P, Liu Q, Wang M, et al. Characteristics of soil fauna community  
517 in the Dongjiao coconut plantation ecosystem in Hainan, China. *Acta Ecologica Sinica*.  
518 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).  
519 3. Singh R, Ong-Abdullah M, Low ET, Manaf MA, Rosli R, Nookiah R, et al. Oil palm genome  
520 sequence reveals divergence of interfertile species in Old and New worlds. *Nature*.  
521 2013;500(7462):335-9. doi:10.1038/nature12309.  
522 4. Al-Dous EK, George B, Al-Mahmoud ME, Al-Jaber MY, Wang H, Salameh YM, et al. De novo  
523 genome sequencing and comparative genomics of date palm (*Phoenix dactylifera*). *Nature*  
524 *biotechnology*. 2011;29(6):521-7. doi:10.1038/nbt.1860.  
525 5. Al-Mssallem IS, Hu S, Zhang X, Lin Q, Liu W, Tan J, et al. Genome sequence of the date palm  
526 *Phoenix dactylifera* L. *Nature communications*. 2013;4:2274. doi:10.1038/ncomms3274.  
527 6. Murray MG and Thompson WF. Rapid isolation of high molecular weight plant DNA. *Nucleic*  
528 *acids research*. 1980;8(19):4321-5.  
529 7. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, et al. SOAPdenovo2: an empirically improved  
530 memory-efficient short-read de novo assembler. *Gigascience*. 2012;1(1):18.  
531 doi:10.1186/2047-217X-1-18.  
532 8. Fan H, Xiao Y, Yang Y, Xia W, Mason AS, Xia Z, et al. RNA-Seq analysis of *Cocos nucifera*:  
533 transcriptome sequencing and de novo assembly for subsequent functional genomics  
534 approaches. *PloS one*. 2013;8(3):e59997. doi:10.1371/journal.pone.0059997.  
535 9. Kent WJ. BLAT—the BLAST-like alignment tool. *Genome Research*. 2002;12(4):656-64.  
536 doi:10.1101/gr.229202. .  
537 10. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV and Zdobnov EM. BUSCO: assessing  
538 genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*  
539 (Oxford, England). 2015;31(19):3210-2. doi:10.1093/bioinformatics/btv351.  
540 11. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O and Walichiewicz J. Repbase  
541 Update, a database of eukaryotic repetitive elements. *Cytogenetic and genome research*.  
542 2005;110(1-4):462-7. doi:10.1159/000084979.  
543 12. Tarailo-Graovac M and Chen N. Using RepeatMasker to identify repetitive elements in  
544 genomic sequences. *Current protocols in bioinformatics*. 2009;Chapter 4:Unit 4.10.

Formatted: Font: Times New Roman, 9 pt

Formatted: Font: Times New Roman

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

doi:10.1002/0471250953.bi0410s25.

546 13. Xu Z and Wang H. LTR\_FINDER: an efficient tool for the prediction of full-length LTR  
547 retrotransposons. *Nucleic acids research*. 2007;35(Web Server issue):W265-8.  
548 doi:10.1093/nar/gkm286.

549 14. Benson G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic acids*  
550 *research*. 1999;27(2):573-80.

551 15. The Arabidopsis Genome Initiative, Analysis of the genome sequence of the flowering plant  
552 *Arabidopsis thaliana*. *Nature*. 2000;408(6814):796-815.  
553 doi:[http://www.nature.com/nature/journal/v408/n6814/supinfo/408796a0\\_S1.html](http://www.nature.com/nature/journal/v408/n6814/supinfo/408796a0_S1.html).

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

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

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

563 19. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and  
564 PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*.  
565 1997;25(17):3389-402.

566 20. Birney E, Clamp M and Durbin R. GeneWise and Genomewise. *Genome Research*.  
567 2004;14(5):988-95. doi:10.1101/gr.1865504.

568 21. Stanke M, Keller O, Gunduz I, Hayes A, Waack S and Morgenstern B. AUGUSTUS: ab initio  
569 prediction of alternative transcripts. *Nucleic acids research*. 2006;34(Web Server  
570 issue):W435-9. doi:10.1093/nar/gkl200.

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

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

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

577 25. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript  
578 assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform  
579 switching during cell differentiation. *Nature biotechnology*. 2010;28(5):511-5.  
580 doi:10.1038/nbt.1621.

581 26. Majoros WH, Pertea M and Salzberg SL. TigrScan and GlimmerHMM: two open source ab  
582 initio eukaryotic gene-finders. *Bioinformatics (Oxford, England)*. 2004;20(16):2878-9.  
583 doi:10.1093/bioinformatics/bth315.

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

586 28. Bairoch A and Apweiler R. The SWISS-PROT protein sequence database and its supplement  
587 TrEMBL in 2000. *Nucleic acids research*. 2000;28(1):45-8.

588 29. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale

Formatted: Font: Times New Roman, 9 pt  
Formatted: Font: Times New Roman

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

589 protein function classification. *Bioinformatics* (Oxford, England). 2014;30(9):1236-40.  
590 doi:10.1093/bioinformatics/btu031.

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

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

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

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

599 34. Mi H, Huang X, Muruganujan A, Tang H, Mills C, Kang D, et al. PANTHER version 11:  
600 expanded annotation data from Gene Ontology and Reactome pathways, and data analysis  
601 tool enhancements. *Nucleic acids research*. 2017;45(Database issue):D183-D9.  
602 doi:10.1093/nar/gkw1138.

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

607 36. Wilson D, Pethica R, Zhou Y, Talbot C, Vogel C, Madera M, et al. SUPERFAMILY--sophisticated  
608 comparative genomics, data mining, visualization and phylogeny. *Nucleic acids research*.  
609 2009;37(Database issue):D380-6. doi:10.1093/nar/gkn762.

610 37. Burge S, Kelly E, Lonsdale D, Mutowo-Muellenet P, McAnulla C, Mitchell A, et al. Manual GO  
611 annotation of predictive protein signatures: the InterPro approach to GO curation. *Database*  
612 : the journal of biological databases and curation. 2012;2012:bar068.  
613 doi:10.1093/database/bar068.

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

616 39. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput.  
617 *Nucleic acids research*. 2004;32(5):1792-7. doi:10.1093/nar/gkh340.

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

621 41. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular biology and*  
622 *evolution*. 2007;24(8):1586-91. doi:10.1093/molbev/msm088.

623 42. Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, et al. The genome of  
624 black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* (New York, NY).  
625 2006;313(5793):1596-604. doi:10.1126/science.1128691.

626 43. Shi H, Lee BH, Wu SJ and Zhu JK. Overexpression of a plasma membrane Na<sup>+</sup>/H<sup>+</sup> antiporter  
627 gene improves salt tolerance in *Arabidopsis thaliana*. *Nature biotechnology*.  
628 2003;21(1):81-5. doi:10.1038/nbt766.

629 44. Brini F, Hanin M, Mezghani I, Berkowitz GA and Masmoudi K. Overexpression of wheat  
630 Na<sup>+</sup>/H<sup>+</sup> antiporter TNHX1 and H<sup>+</sup>-pyrophosphatase TVP1 improve salt- and drought-stress  
631 tolerance in *Arabidopsis thaliana* plants. *Journal of experimental botany*. 2007;58(2):301-8.  
632 doi:10.1093/jxb/erl251.

Formatted: Font: Times New Roman, 9 pt  
Formatted: Font: Times New Roman

633 45. Yong JW, Ge L, Ng YF and Tan SN. The chemical composition and biological properties of  
 634 coconut (*Cocos nucifera* L.) water. *Molecules* (Basel, Switzerland). 2009;14(12):5144-64.  
 635 doi:10.3390/molecules14125144.

639 **Tables**

640 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)

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

644 Table 2 Comparison analysis of genome sizes, assembly and annotation of four palmae species, including  
 645 coconut, *Phoenix dactylifera* (PDK30 and DPV01, two different versions), *Elaeis guineensis* (EG), and *Elaeis*  
 646 *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

647 Note: Coconut: *Cocos nucifera* (Hai-nan Tall); PDK30: *Phoenix dactylifera* (PDK30); DPV01: *Phoenix*  
 648 *dactylifera* (DPV01); EG: *Elaeis guineensis* (American-African oil palm E5 build); EO *Elaeis oleifera* (American  
 649 oil palm, O8-build); The-TEs results were as obtained using the same pipeline as for the with the Coconut

Formatted: Left

Formatted: Font: Times New Roman, 9 pt

Formatted: Font: Times New Roman

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

650 [genome](#)  
651  
652  
653  
654  
655  
656  
657  
658

Table 3 The gene coverage of *Cocos nucifera* [by-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

659

660 Table 4\_The comparative analysis of assembly results of five\_palm species with BUSCO\_software, including  
661 coconut, *Phoenix dactylifera* (PDK30 and DPV01, two varieties ), *Elaeis guineensis* (EG), and *Elaeis oleifera*  
662 (EO)

Formatted: Left

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

663 Note: Coconut: *Cocos nucifera* (the Hainan Tall); PDK30: *Phoenix dactylifera* (PDK30); DPV01:*Phoenix*  
664 *dactylifera* (DPV01); EG: *Elaeis guineensis*(~~African~~american oil palm E5 build); EO *Elaeis oleifera*\_(American  
665 oil palm, O8-build);

666

667 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
Total	385,037,442	546,965,774	1,552,582,881	1,602,630,396	72.75

Formatted: Font: Times New Roman, 9 pt

Formatted: Font: Times New Roman

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

668 Note: Repbase TEs means RepeatMask against Repbase; Protein TEs means RepeatProteinMask result against  
669 Repbase protein; *De novo* TEs means RepeatMask against the *de novo* library; Combined TEs: ~~means~~ the  
670 combined results of these three steps.

671  
672  
673  
674  
675  
676

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

677 Note: Coconut: *Cocos nucifera* (the Hai-nan Tall); PDK30: *Phoenix dactylifera* (PDK30); DPV01: *Phoenix*  
678 *dactylifera* (DPV01); EG: *Elaeis guineensis* (AfricanAmerican oil palm E5 build); The gene of *Elaeis oleifera*,  
679 (American oil palm, O8-build) was missing, not attained from the public database;

680  
681

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

682  
683  
684  
685

Formatted: Left

Formatted: Font: Times New Roman, 9 pt

Formatted: Font: Times New Roman



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

686  
687  
688  
689  
690  
691  
692  
693  
694  
695  
696  
697  
698  
699  
700  
701  
702  
703  
704  
705  
706  
707  
708  
709  
710

**Figure legends**

Figure 1 Morphological characteristic of coconut tree (A), spica (B), female flower (C), Male flower (D), –coconut nut (E), coconut nut without skin (F), and –vertical section of coconut nut (G).

Figure 2 Kmer analysis of the coconut genome.

Figure 3 The protocol for integrating GLEAN and [Transcriptome-transcriptome](#) data.

[Figure 4 Groups of orthologues shared among the angiosperms \*Cocos nucifera\* \(Coconut\), \*Elaeis guineensis\* \(Oil palm\), \*Phoenix dactylifera\* \(Date palm\), \*Sorghum bicolor\* \(Sorghum\), \*Musa acuminata\* \(Banana\) and \*Arabidopsis thaliana\* \(Arabidopsis\). Venn diagram generated by <http://www.interactivenn.net/>.](#)

[Figure 5. Estimation of divergence time. The blue numbers on the nodes are the divergence time from present \(million years ago, Mya\), the red nodes indicated the previously published calibration times.](#)

[Figure 6. Phylogenetic tree of antiporter genes from \*C. nucifera\* and \*Arabidopsis thaliana\*. Every cluster was indicated with a different colored arc line arc. The potential function of every cluster was indicated with the function groups found in \*Arabidopsis thaliana\*. Colored stars indicate antiporter genes of \*C. nucifera\*.](#)

[Figure 7. Phylogenetic tree of ion channel genes from \*C. nucifera\* and \*Arabidopsis thaliana\*. Every cluster was indicated with different colored arc line arc. The potential function of every cluster was indicated with the function groups found in \*Arabidopsis thaliana\*. Colored stars indicate ion channel genes of \*C. nucifera\*.](#)

Formatted: Font: Times New Roman, 9 pt  
Formatted: Font: Times New Roman

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

711 Figure 4 Groups of orthologues shared among the angiosperms *Cocos nucifera* (Coconut), *Elaeis*  
712 *guineensis* (Oil palm), *Phoenix dactylifera* (Date palm), *Sorghum bicolor* (Sorghum), *Musa*  
713 *acuminata* (Banana) and *Arabidopsis thaliana* (Arabidopsis). Venn diagram generated by  
714 <http://www.interactivenn.net/>—

715 Figure 5. Estimation of divergence time. The blue numbers on the nodes are the divergence time from  
716 present (million years ago, Mya), the red node indicated the previously published calibration times.

717 Figure 6. Phylogenetic tree of antiporter genes from *C. nucifera* and *Arabidopsis thaliana*. Every  
718 cluster was indicated with different colored arc line. The potential function of every cluster was  
719 indicated with the function groups found in *Arabidopsis thaliana*. Colored stars indicate antiporter  
720 genes of *C. nucifera*.

721 Figure 7. Phylogenetic tree of ion channel genes from *C. nucifera* and *Arabidopsis thaliana*. Every  
722 cluster was indicated with different colored arc line. The potential function of every cluster was  
723 indicated with the function groups found in *Arabidopsis thaliana*. Colored stars indicate ion channel  
724 genes of *C. nucifera*.

725

726

727

728

729

730

731 [Additional files](#)

732

733 [Additional file 1 Identification and characterization of antiporter genes in the genome of \*Cocos\*](#)

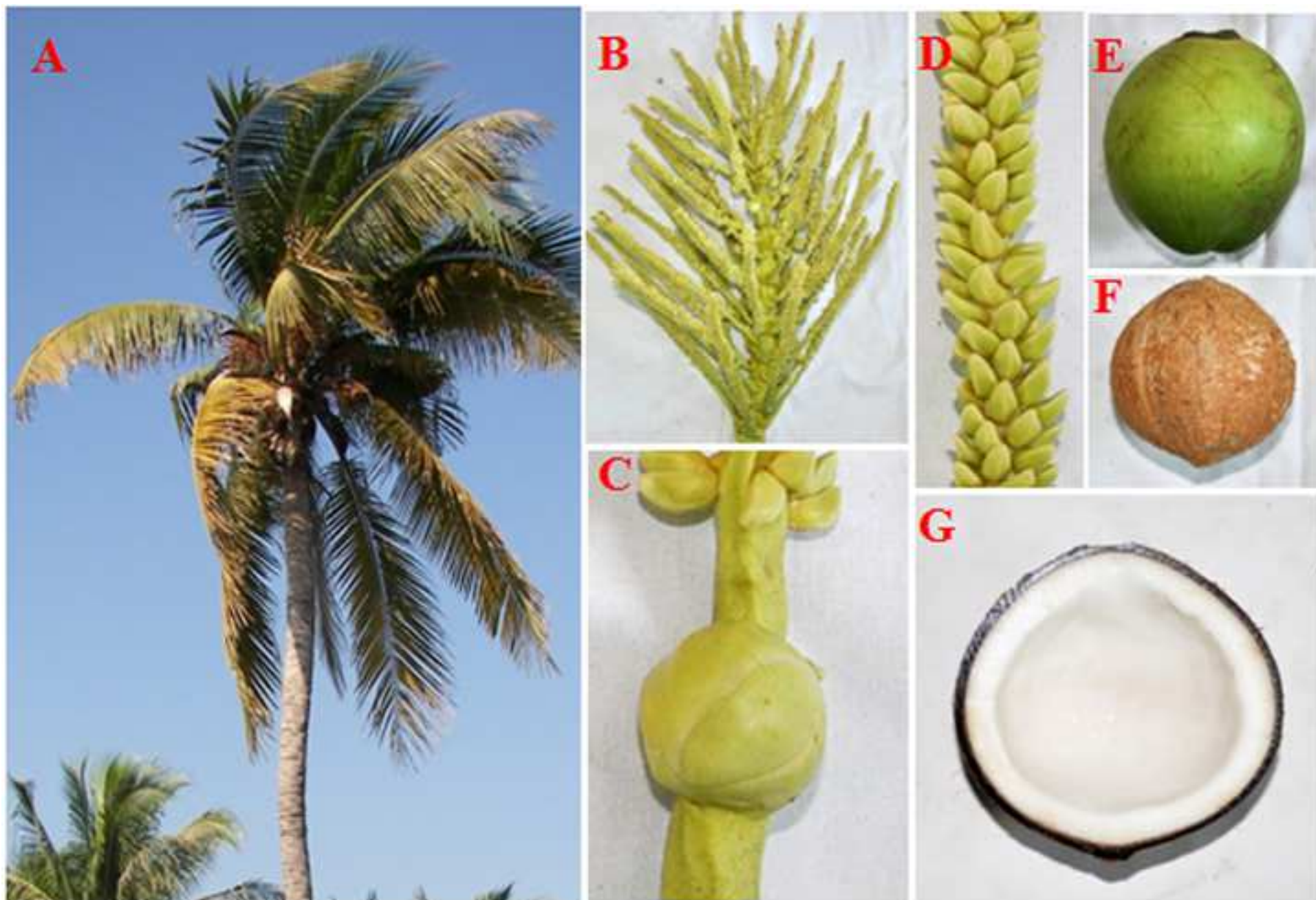
734 [nucifera](#)

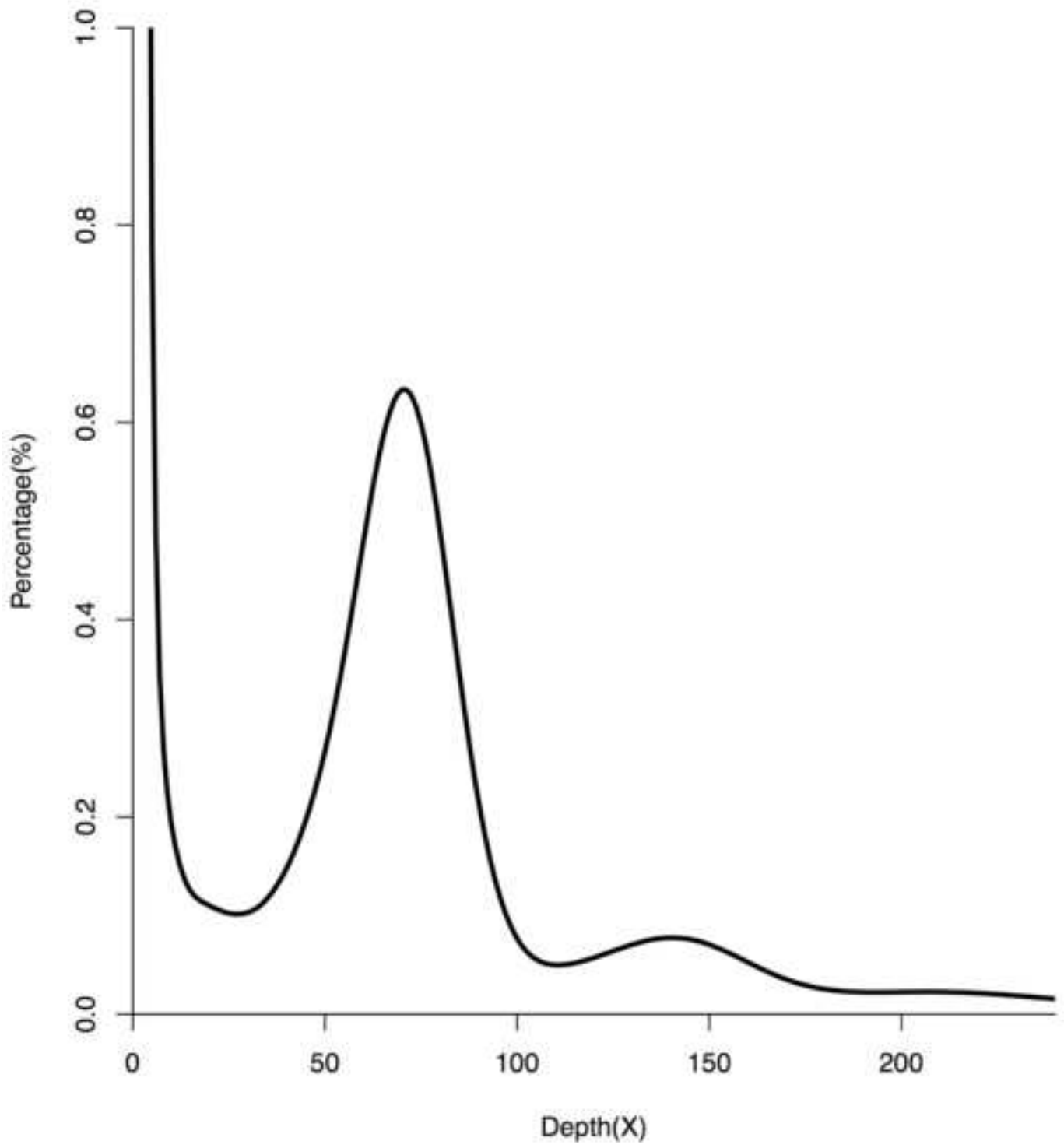
735

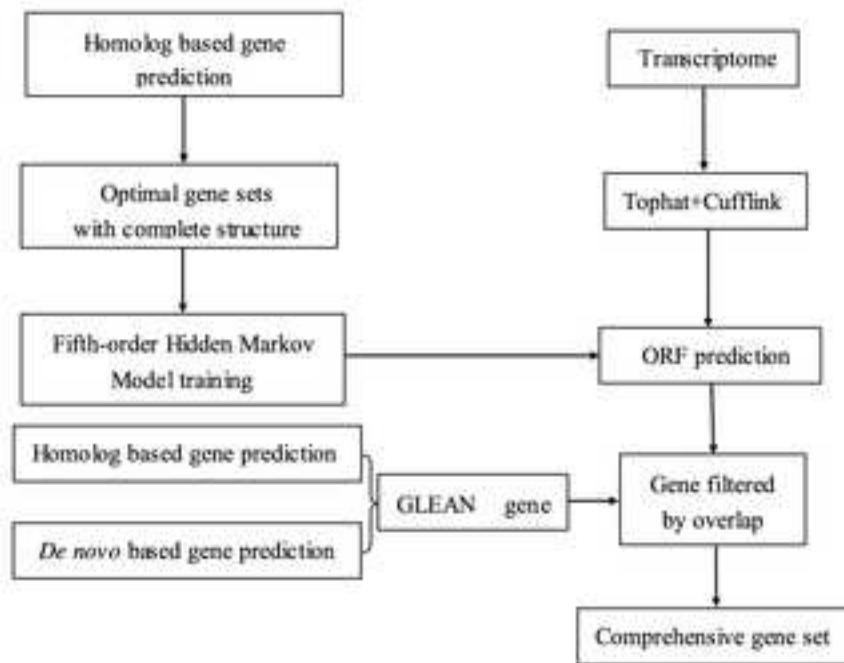
736 [Additional file 2 Identification and characterization of ion channel genes in the genome](#)

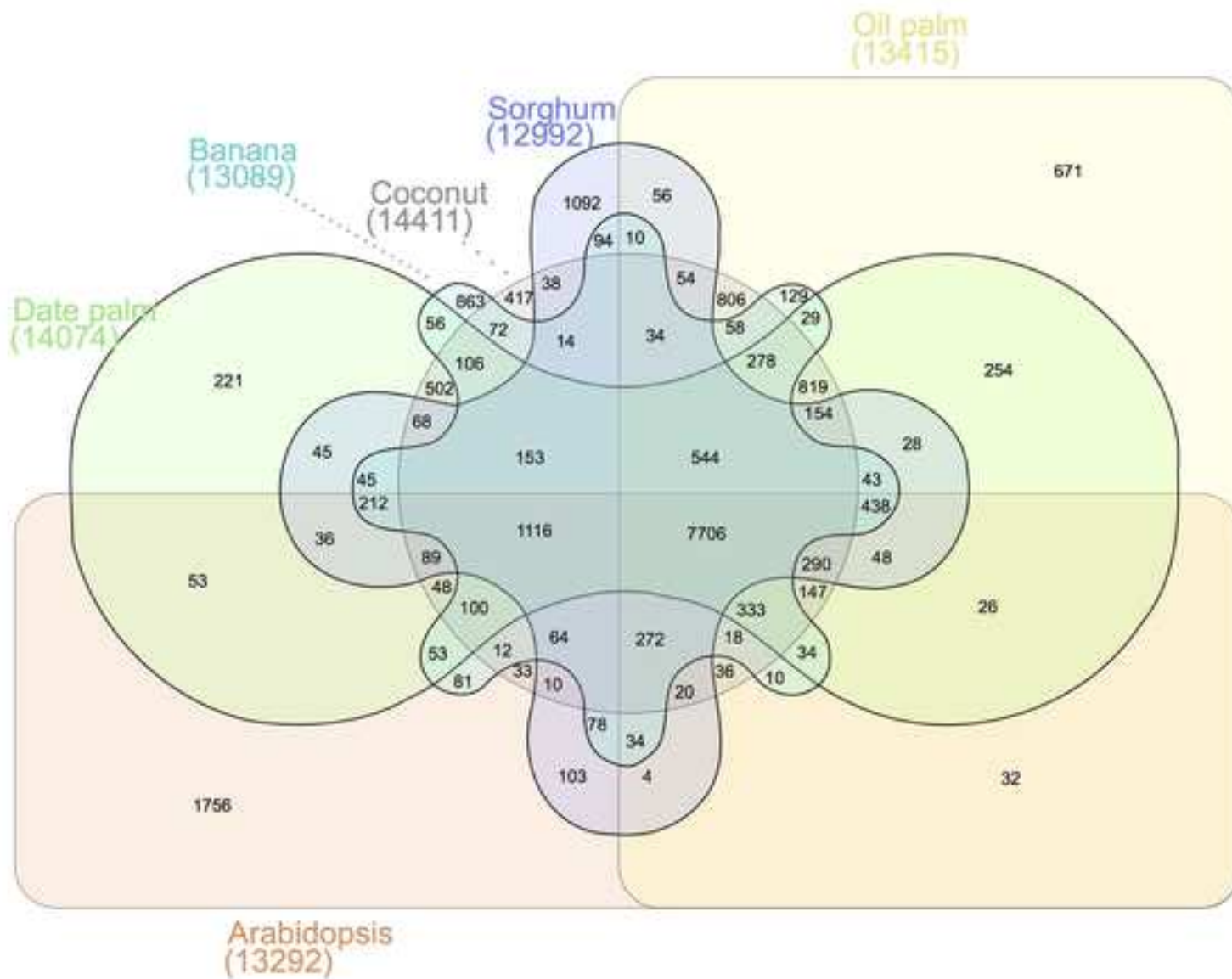
737 [of \*Cocos nucifera\*](#)

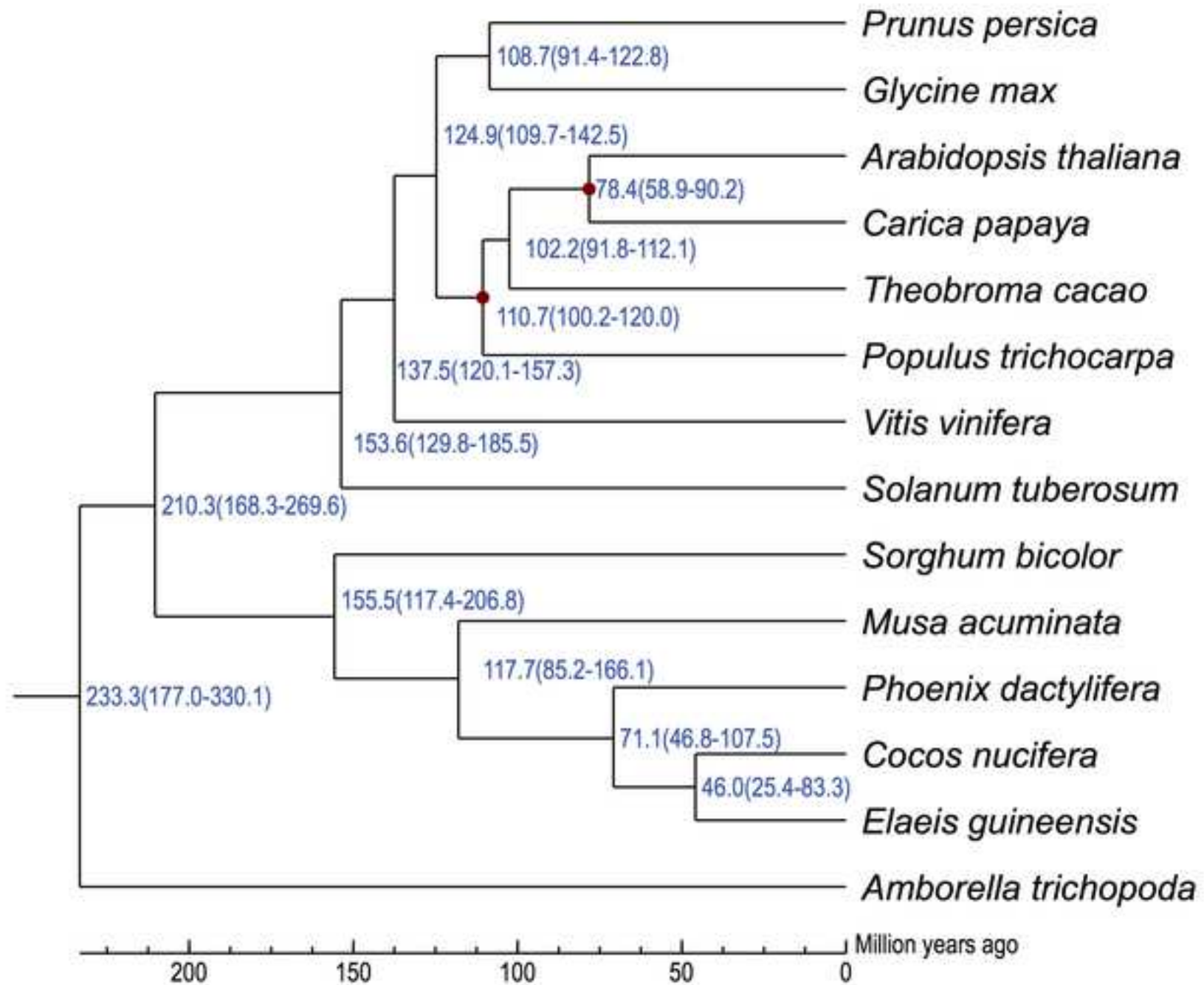
Formatted: Font: Times New Roman, 9 pt  
Formatted: Font: Times New Roman

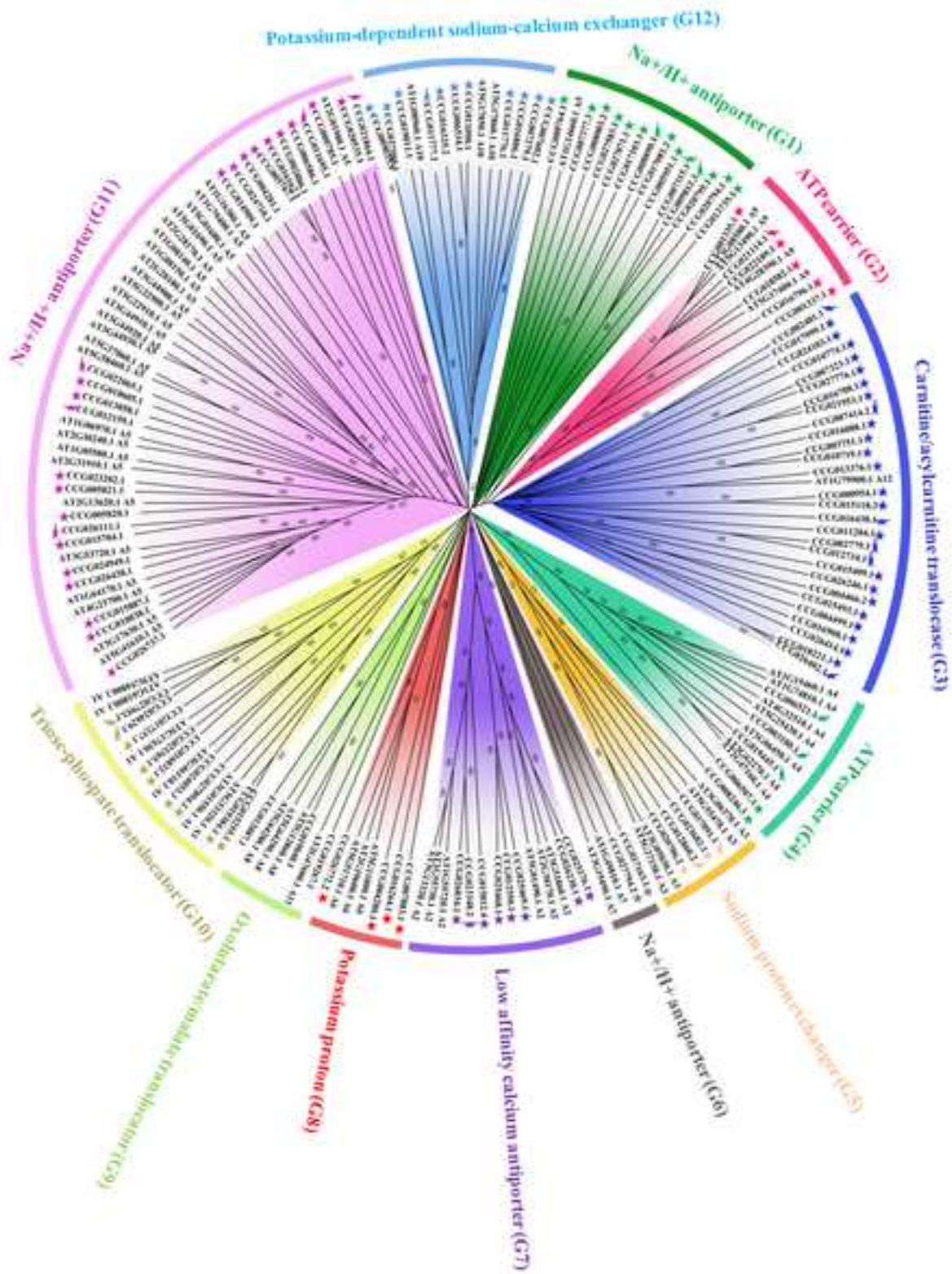




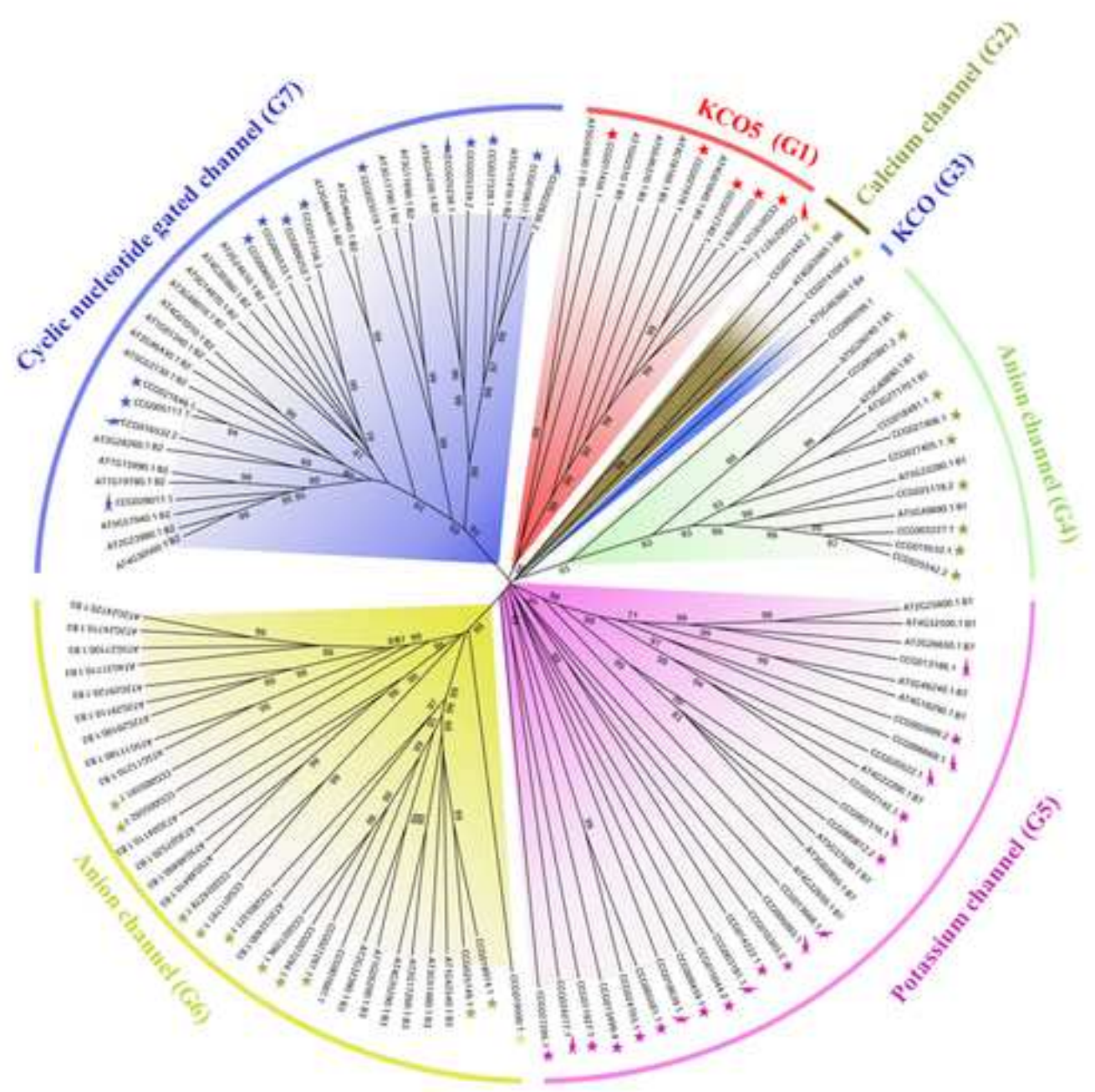


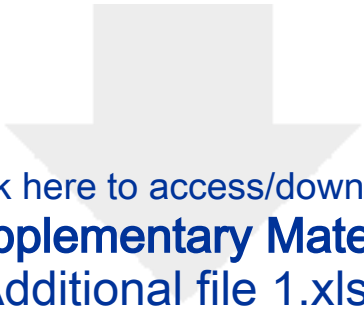







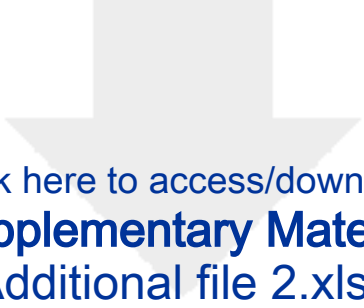




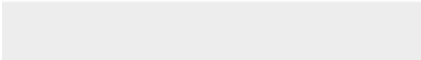



Click here to access/download  
**Supplementary Material**  
Additional file 1.xlsx





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



## Response to editor and reviewers

Dear editor and reviewers

Thank you very much for your crucial comments for our manuscript entitled “The genome draft of the Coconut (*Cocos nucifera*)” (GIGA-D-17-00038). We have made a thorough revision to the ms based on all of comments from editor and reviewers. Each comments raised by the reviewers had been carefully answered in the response sheet. We hope the revised version can meet the requirement of “GigaScience”

Sincerely yours,

Yaodong Yang

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

Response to editor and reviewers,

Reviewer 1

1. Line 40: in 93 countries -> the introduction (line 70) say 89 countries

>>>Response: Thank you for your suggestion. We have re-checked the document reported by Batugal et al., 2005. The corresponding revision has been done in the Introduction part of the revised manuscript.

2. 11 million ha ->the introduction (line 72) says 12 million ha

>>>Response: Thank you for your suggestion; we have re-checked the plant area of coconut in the website of Food and Agriculture Organization of the United Nations (<http://www.fao.org/faostat/en/>). The corresponding revision has been done in the Abstract part of the revised manuscript.

3. Hinders progress in genetic breeding. Do you mean ‘marker assisted breeding’ or ‘genomic assisted breeding’?

>>>Response: Thank you for your suggestion; we meant to say 'conventional breeding'. Revisions have been made in the Abstract part of the revised manuscript to make our opinions clearer.

4. Genetic improvement is slow. Do you mean trait improvement with marker or genetic assisted

>>>Response: We meant to say the improvement made by 'conventional breeding' is slow. The corresponding revision has been done in the revised manuscript.

5. Line 48: The coverage does not add up. 714.67 Gb on a 2.42 Gb genome is 295× coverage. In any case, only the coverage of the cleaned reads should be shown (177×)

>>>Response: Thank you for your suggestion; in revised manuscript, only the cleaned reads were used for the coverage depth analysis and the coverage is 173.32× read depth.

6. Line 54: Do you mean 41,166 genes

>>>Response: Thank you for your suggestion; we have re-checked the annotated gene number for date palm based on the document reported by AI-Mssallem et al., 2013 and 41 660 genes were annotated. The corresponding revisions have been made in the Abstract part of revised manuscript.

7. Line 60: space missing between facilitating and future

>>>Response: Thank you for your suggestion, a space has been added between facilitating and future.

8. Line 61: should be 'molecular assisted breeding'

>>>Response: Thank you for your suggestion, corresponding revisions have been done in the Abstract part of revised version.

9. Line 78: '...wide range to environment...' -> unclear, should be explained. Also 'environment'

>>>Response: Some sentences have been added to the revised manuscript for explaining '...wide range to environment...' in Line 240– Line 242|Page 3.

10. Line 78: '...especially for high tolerance to high salt density.', please clarify

>>>Response: Coconut palm can disseminate through ocean currents: floating nuts sprout and

grow naturally upon washing up on beaches. The ability to adapt to a high salt environment is closely related to this dissemination feature and to these natural growth conditions. Corresponding revision has been done in Line 243– Line 244|Page 3 of revised manuscript.

11. Line 80: ‘...making it possible to understand its adaptation to high salinity.’ You do not investigate this, you should change the statement to something milder such as: ‘This study forms the basis for future research investigating the coconuts tolerance to salt stress’

>>>Response: Thank you for your suggestion, We also present the genome sequence of HAT coconut and added an analysis of the antiporter and ion channel gene families, relevant to salinity tolerance, into the revised version. Corresponding revision had been added into in Line 237– Line 238|Page 3.

12. Line 82: provide references. The way this sentence reads at the moment, make it seem like you are also reporting those genome sequence.

>>>Response: The corresponding references have been added into Line 423|Page 4 of revised manuscript.

13. Line 92: space between ‘Illumina’, ‘Hiseq2000’ and ‘sequencer’

>>>Response: Two spaces had been added into between Illumina, Hiseq2000 and sequencer in Line 436|Page 4 of revised manuscript.

14. Line129: The data shows that you have higher coverage and a longer N50, it does not show that the assembly is of better quality.

>>>Response: Thank you for your suggestion, the sentence has been replace by other sentence: “The comparative results of the BUSCO estimation in coconut and in the four other palm genome sequences indicates that the smallest fraction of missing genes as predicted by BUSCO was found in the coconut genome assmebly”, in Line 724 – Line 726|Page 6 of revised version.

15. Line 131: ‘tissues’, not ‘issues’

>>>Response: Thank you for your suggestion, corresponding revisions has been done in the revised version.

16. Line134: table 4 and 5 are mixed up

>>>Response: We repeatedly checked Table 4 and 5. Corresponding revisions has been done in revised manuscript.

17. Line 165: BLAST not BLSAT

>>>Response: Thank you for your suggestion, 'BLSAT' had been modified in revised manuscript.

18. Line 175 (and others): keep a space between numbers and units, consistently.

>>>Response: we re-checked all numbers and units throughout the manuscript. All needed spaces have been added between numbers and units.

19. Line195: Change start of sentence (e.g. 'After the above described steps...')

>>> Response: Thank you for your suggestion, corresponding revision has been done in Line 970|Page 8 of the revised manuscript.

20. Line 196: should read: 'than the predicted gene markers..'

>>>Response: Thank you for your suggestion, corresponding revision has been done in Line 971 | Page 8 of revised version.

21. Line203: space between 'by' and 'sequence'

>>>Response: Thank you for your suggestion, a space had been added between by and sequence

22. Line211: after ref 38, just one dot

>>>Response: Thank you for your suggestion, the ref 38 and dot has been deleted in revised version.

23. Line 219: remove space between 'mapping' and ','

>>>Response: Thank you for your suggestion, the space has been deleted between 'mapping' and ','.

24. References: need a lot of editing to uniform

>>>Response: All references of the manuscript have been reviewed and edited based on the author guideline of "Gigascience" in the revised manuscript.

25. Tables: Headers are unclear and many abbreviations within tables are not explained

>>>Response: Thank you for your suggestion, revisions have been done for the table headers. Meanwhile, the abbreviations have been explained and replaced with corresponding full name.

26. What is the difference between Table 4 and Table 7? Both show BUSCO assessments of palm species. Clarify both in tables and in the text.

>>>Response: Thank you for your suggestion, Table 7 has been changed into Table 6 in the revised version. Table 4 referred to the comparative analysis of the assembled genome sequences for four palm species using BUSCO software, while Table 6 referred to the comparative analysis of the predicted gene from the four palm species using BUSCO software. Revisions have been done to make Table 4 and Table 6 legends more clearly in “Table” part of revised version.

27. Figure legend: Figure 1 does not contain any morphological characteristics; they are photographs of coconut plants.

>>>Response: Figure 1 had been substantially revised in the revised version.

#### Reviewer 2

1. My only major concern about the manuscript is that the written style is not ready for publication. There are many type and grammatical mistakes all over the main text, figure captions and table legends. The manuscript needs some extensive copy editing to be published.

>>>Response: Thank you for your suggestion, the manuscript has been reviewed and edited throughout the manuscript by the native experts (Annaliese Mason, Baudouin Luc and Amjad Iqbal).

#### Reviewer 3

1. Homologous gene families using a larger set of genomes would allow a gain-/loss analysis (check the *Zostera* (seagrass) genome paper Figure 1a for a recent example), some venn diagrams based on this showing how many gene are shared with close relative (e.g. *Elaeis*), other monocots (e.g. rice) and dicots (e.g. Arabidopsis) could also be generated based on this (e.g. orchid genome paper figure 1a). Asynteny/collinearity analysis is usually included, often combined with a Ks analysis (see the orchid genome paper Figure 2, *Zostera* genome paper Figure 2).

>>>Response: Thank you for your suggestion, we added venn diagrams between different species and analyzed the divergence time between different species into Line 990| Page 8 - Line 1270 | Page 10 of the revised version. Meanwhile, we identified and characterized antiporter and ion channel gene family in Line 1271 | Page 10 – Line 1578 | Page 11 of revised manuscript.

2. No case study is included, I feel there should be at least one (though as the paper is submitted as a data note the journal might not require one). The authors are the first ones to have a glimpse at the genome of this species. I would make sense to check a few relevant gene families (coconut are clearly very different from seeds of other monocots, so seed related gene families would be likely candidates for a more in depth study)

>>>Response: Thank you for your suggestion. It is known that coconut palm can disseminate



through ocean currents: floating nuts sprout and grow naturally upon washing up on beaches. The ability to adapt to a high salt environment is closely related to this dissemination feature and to these natural growth conditions. In the revised manuscript, we identified antiporter and ion channel genes in the genome of *Cocos nucifera*, some of which had been validated to be associated with salt stress in Arabidopsis. In the gene expansions analysis, some gene families showed significant expansion in compared to Arabidopsis, including Na<sup>+</sup>/H<sup>+</sup> antiporter family, Carnitine/acylcarnitine translocase family, Potassium-dependent sodium antiporter, and potassium channel. The expansion of Na<sup>+</sup>/H<sup>+</sup> antiporter family and Potassium-dependent sodium antiporter may be associated with coconut salt tolerance. The expansion of carnitine/acylcarnitine translocase family may be associated with the accumulation of fatty acid in coconut pulp. At last, the expansion of potassium channel may be associated with the accumulation of potassium ion in coconut water. Corresponding revision had been added into Line Line 1271 | Page 10 – Line 1578 | Page 11 of revised manuscript.

3. For non-bioinformaticians a supplemental website which offers a BLAST interface would certainly be welcome.

>>>Response: we have uploaded coconut genome raw data into Sequence Read Archive (SRA) of the National Center for Biotechnology Information. The assembled and annotated data were uploaded into GigaDB database. Meanwhile, the assembled and annotated data have been uploaded into pirate website for blast analysis and genome browse. However, currently, this website is not available for all people. The website will be available after further website improvement and paper publication

4. Line 128 -129: The N50 by itself is not a direct measure for the quality of the assembly. Avoid over-interpretation.

>>>Response: Thank you for your suggestion, the sentence has been replace by other sentence: “The comparative results of the BUSCO estimation in coconut and in the four other palm genome sequences indicates that the smallest fraction of missing genes as predicted by BUSCO was found in the coconut genome assmebly”, in Line 724 – Line 726|Page 6 of revised version.

5. Line 54 and abstract: (DVP01, 4166) -> the number of genes for the date palm genome is incorrect. In table 2 the authors report 41, 660 !

>>>Response: Thank you for your suggestion; we have re-checked the annotated gene number for datepalm based on the document reported by AI-Mssallem et al., 2013 and 41 660 genes were annotated. The corresponding revisions have been made in the Abstract part of revised manuscript.

6. Line 60: facilitating future: missing space Line 78-79

>>>Response: Thank you for your suggestion, a space has been added into between facilitating and future.

7. Line 78-79: For high tolerant to high salt density: revise grammar

>>>Response: Thank you for your suggestion, revisions have been done in Line 240– Line 242 | Page 3 of revised manuscript

8. Line 80: ...present Hainan Tall... -> ...present the Hainan Tall

>>>Response: Thank you for your suggestion, the sentence has been rewrite and the usage of the phrase “Hainan Tall” has been carefully checked throughout the revised manuscript.

9. Line 82: ...about genome

>>>Response: Revisions have been done in revised manuscript.

10. Line 88 (and other places): ...pair end... -> ...paired end...

>>>Response: Thank you for your suggestion, all ‘pair end’ has been modified into ‘paired end’ throughout the revised manuscript.

11. Line 96: ...removed by using ... -> ...removed using...

>>>Response: Thank you for your suggestion, corresponding revision had been done in revised manuscript

12. Line 116: ...SOAPdenovo2 map... -> ...SOAPdenovo2 maps...

>>>Response: Thank you for your suggestion, corresponding revision had been done in Line 572 | Page 5 of revised manuscript.

13. Line 132: ...reported in previous Fan’s research ...: incorrect grammar should be revised (as previously reported by Fan et al...).

>>>Response: Thank you for your suggestion, corresponding revision had been done in Line 598 | Page 5 in revised manuscript

14. Line 181: Previous Fan’s research: revise grammar

>>>Response: ‘Previous Fan’s research’ had been modified into ‘as previously reported by Fan et al.’ in Line 874 | Page 7 of revised manuscript

15. Line 192: ... a diagrammic pipeline is showed...: revise grammar

>>>Response: Thank you for your suggestion, corresponding revisions had been done in Line

968|Page 8 revised version.

16. Line 199: ...completely... -> complete

>>>Response: Thank you for your suggestion, corresponding revision has been done in Line 973 | Page 8 of revised manuscript.

17. Line 203 -204: In sequence similarity step: revise

>>>Response: 'In sequence similarity step' has been modified into 'Firstly' in Line 979|Page 8 of the revised manuscript.

18. Line 232-233: Font is suddenly somewhat bigger

>>>Response: Thank you for your suggestion, corresponding revision have been done in "Funding" part of revised manuscript.