Complete Sequence and Comparative Analysis of the Chloroplast Genome of Coconut Palm (Cocos nucifera)

Ya-Yi Huang, Antonius J. M. Matzke, Marjori Matzke*
Institute of Plant and Microbial Biology, Academia Sinica, Taipei, Taiwan

Abstract

Coconut, a member of the palm family (Arecales), is one of the most economically important trees used by mankind. Despite its diverse morphological, coconut is recognized taxonomically as only a single species (Cocos nucifera L.). There are two major coconut varieties, tall and dwarf, the latter of which displays traits resulting from selection by humans. We report here the complete chloroplast (cp) genome of a dwarf coconut plant, and describe the gene content and organization, inverted repeat fluctuations, repeated sequence structure, and occurrence of RNA editing. Phylogenetic relationships of monocots were inferred based on 47 chloroplast protein-coding genes. Potential nodes for events of gene duplication and pseudogenization related to inverted repeat fluctuation were mapped onto the tree using parsimony criteria. We compare our findings with those from other palm species for which complete cp genome sequences are available.

Citation: Huang Y-Y, Matzke AJM, Matzke M (2013) Complete Sequence and Comparative Analysis of the Chloroplast Genome of Coconut Palm (Cocos nucifera). PLoS ONE 8(8): e74736. doi:10.1371/journal.pone.0074736

Editor: Hector Candela, Universidad Miguel Hernández de Elche, Spain
Received June 25, 2013; Accepted August 6, 2013; Published August 30, 2013
Copyright: © 2013 Huang et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work was funded by Academia Sinica (http://www.sinica.edu.tw/main_e.shtml). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: marjorimatzke@gate.sinica.edu.tw

Introduction

Chloroplasts (cp) are cell organelles that carry out photosynthesis, thus converting light energy into chemical energy in green plants and algae. Chloroplasts contain their own genome, which in flowering plants usually consists of a circular double-stranded DNA molecule ranging from 120 to 160 kb in length [1]. The cp genome is divided into four parts comprising a large single copy region (LSC) and a small single copy region (SSC), which are separated by a pair of inverted repeats (IRs). Cp genomes typically encode four rRNAs, around 30 tRNAs and up to 80 unique proteins [2–4].

With the advent of high-throughput sequencing technologies and their use in obtaining complete plastid genomes [5,6], the number of fully sequenced cp genomes has increased rapidly. To date, the Complete Organelle Genome Sequences Database (http://amoebidia.bcm.umontreal.ca/pg-gobase/complete_genome/ogmp.html) lists 324 complete cp genome sequences spanning 268 distinct organisms. The complete cp genome sequences include date palm (Phoenix dactylifera L.) and oil palm (Elaeis guineensis Jacq.). Both are members of the palm family (Arecales), which is the most economically important family of plants after the grasses and legumes [7]. Complete sequence information on cp genomes from three additional palms - Calamus caryotoides, Pseudophoenix sargentii, Bismarkia nobilis – has recently been deposited in GenBank [8]. However, the complete cp genome sequence of coconut palm (Cocos nucifera L.), which is a universal symbol of the tropics and equally important as oil palm [7], has not yet been reported.

Coconut is one of the most important crops in tropical zones where it is a source of food, drink, fuel, medicines and construction material [9]. In addition, coconut oil is used for cooking and for pharmaceutical and industrial applications [10]. Although coconut trees display considerable morphological diversity, they are considered taxonomically a single species (and the only species) within the genus Cocos. Based on stature and breeding, coconut cultivars can be divided into two groups: tall and dwarf [11]. The former typically grows up to 35 to 40 meters and is mainly outcrossing, whereas the latter can only grow up to 25 to 30 meters and usually is selfing. Dwarf coconuts, which are less common than the tall variety, are usually found growing close to humans and have traits that likely result from human selection [10]. Here we report the complete cp genome sequence of a dwarf coconut plant, which is thought to be descended from coconut trees originally imported into Taiwan from Thailand (personal communication from private breeder).

Materials and Methods

Whole genome sequencing and de novo assembly

Fresh young leaf material (ca. 2 g) was collected from a coconut seedling growing under ambient conditions in the greenhouse of Academia Sinica and the genomic DNA (gDNA) was extracted using a modified CTAB protocol [12]. We used the ratio of absorbance at 260 nm and 280 nm (A260/280 = 1.8) to estimate the purity and integrity of the extracted gDNA. High quality DNA (concentration >100 ng/µL; A260/230 >1.7; A260/280 = 1.8–2.0) was sequenced using the Illumina GAIIx platform (YOURGENE BIO SCIENCE Co., New Taipei City, Taiwan). Short reads (70 bp) from paired-end sequencing were trimmed with a 0.05 error probability. The trimmed reads were de novo assembled using CLC Genomic Workbench 6.0.1 (CLC Bio, Aarhus, Denmark). The de Bruijn Graph approach with a k-mer length of 22 bp and a coverage cutoff value of 10X was applied for assembly. The average read length and insert size...
were 151 bp and 340 bp respectively. The assembled contigs shorter than 200 bp were removed from the scaffold while those with coverage larger than 10X were selected for BLAST search against plastid genomes of date palm [2], oil palm [3], and other chloroplast sequences with an e-value cutoff of 10^-2 (199 sequences in total). Gaps between contigs were filled by PCR amplification with specific primers that were designed based on contig sequences or homologous sequence alignments (Table S1). The PCR products were purified with GEL/PCR DNA clean-up kit (Favorgen Biotech Corp.) and then sequenced by conventional Sanger sequencing. The sequencing data along with gene annotation have been submitted to GenBank with an Accession number of KF285453.

### Genome annotation, base composition, repeat structure, and codon usage

Preliminarily gene annotation was carried out through the online program DOGMA [13] and BLAST searches. To verify the exact gene and exon boundaries, we used MUSCLE [14] to align putative gene sequences with their homologues acquired from BLAST searches in GenBank. All tRNA genes were further confirmed through online tRNAscan-SE search server [15].

The online program tandem repeat finder [16] was used to search the locations of repeat sequences (>10 bp in length) with the following set up: (2, 7, 7) for alignment parameters (match, mismatch, indels); 80 for minimum alignment score to report repeat; and maximum period size of 500. Codon usage was calculated for all exons of protein-coding genes (pseudogenes were not calculated). Base composition was calculated by Artemis [17].

### Analysis of RNA editing

Potential RNA editing sites in protein-coding genes of coconut cpDNA were predicted by the online program Predictive RNA Editor for Plants (PREP) suite (http://prep.unl.edu/) [18] with a cutoff value of 0.8. This program contains 35 reference genes for detecting RNA editing sites in plastid genomes. The predicted editing sites were verified by reverse transcription polymerase chain reaction (RT-PCR) experiments. In addition to those genes predicted by the program, we also investigated rpl22, rpl23, rps3, rps7, rps11, rps12, and rpl4 genes, within which RNA editing sites were reported in the cp genome of oil palm [3]. The Plant Total RNA Miniprep Purification Kit (GMbiolab Co., Ltd.) was applied to extract total RNA from leaf of the same seedling used for DNA extraction. The first strand cDNA was synthesized with Quanti-
Tect Reverse Transcription Kit (Qiagen) following the manufacturer’s protocol. Gene specific primers for cDNA amplification were designed based on homologous sequence alignment. Maximum 1 µl of the reaction mixture was used as template for PCR amplification. The PCR products were purified with GEL/PCR DNA clean-up kit (Favorgen Biotech Corp.). Purified PCR products were sequenced using ABI PRISM H3700. A complete primer list is provided in Table S1.

Phylogenetic analysis
Forty seven protein coding genes were extracted from 25 taxa, including *Amborella, Nuphar*, 17 species of monocots, four species of magnoliids, and two species of eudicots. The GenBank accession number of each taxon is provided in Table 1. These taxa were selected because they have complete or nearly complete plastid genomes deposited in GenBank. Nucleotide sequences of each gene were first aligned by MUSCLE [14] through the online server of European Bioinformatics Institute (http://www.ebi.ac.uk).

Figure 1. Coconut chloroplast genome map. Genes shown on the outside of the large circle are transcribed clockwise, while genes shown on the inside are transcribed counterclockwise. Thick lines of the small circle indicate IRs. Genes with intron are marked with “*”. Pseudo genes are marked with “Y”.

doi:10.1371/journal.pone.0074736.g001
uk/Tools/msa/muscle). The aligned sequences were then concatenated through copy and paste in text editor. The statistical method of Maximum Likelihood (ML) and the computer program Garli version 2.0 were applied for phylogenetic reconstruction, with parameters estimated from the data. The GTR substitution model with evolutionary rates among sites evaluated by a discrete gamma distribution was used for tree search. All positions containing gaps or missing data were eliminated. Branch support was evaluated by 1,000 replications of bootstrap (BS) re-sampling.

Results and Discussion

Sequencing and de novo assembly

Illumina sequencing produced 6,413,504 paired-end reads with an average read length of 151 bp and a total base number of 968,439,104. After quality trim, 6,328,120 reads with an average of 145.3 bp and a total base number of 919,475,836 remain. The subsequent de novo assembly and reference-guided blast search resulted in five major contigs separated by five gaps, which were then filled by Sanger sequencing. In addition to gap closure and confirmation of four junction regions (LSC/IR\textsubscript{A}, LSC/IR\textsubscript{B}, SSC/IR\textsubscript{C}, SSC/IR\textsubscript{B}), we also validated the accuracy of our whole genome sequencing by randomly selecting genes/spacers for PCR-based sequencing. Priority was given to long genes (e.g., \textit{ycf1}, \textit{ycf2}, \textit{rpoC1}) or long spacers (between \textit{rpoB} and \textit{psbD}, \textit{ycf2} and \textit{ndhB}, \textit{ndhC} and \textit{trnV-UAC}). A few regions where genes were transcribed from clockwise to counterclockwise (vice versa) were also validated.

Organization of chloroplast genome

Analysis of the data obtained from high-throughput sequencing demonstrated that the cp genome of coconut is a typical quadripartite molecule (Fig. 1) within which a pair of inverted repeats (IRs) is separated by a large single copy region (LSC) and a small single copy region (SSC). The genome is 154,731 bp in length (IR\textsubscript{A} = 53,110 bp; LSC = 84,230 bp; SSC = 17,391 bp) and is predicted to encode 130 genes and four pseudogenes. The former includes 84 protein-coding genes, 36 tRNA genes, and

| Amino acid | Codon | No | RSCU | tRNA | Amino acid | Codon | No | RSCU | tRNA |
|------------|-------|----|------|------|------------|-------|----|------|------|
| Phe        | UUU   | 906| 1.23 |      | Ala        | GCA   | 383| 0.59 |      |
|            | UUC   | 564| 0.77 | \textit{trnF-GAA} | GAC   | 202| 0.31 |      |
|             |       |    |      |      |             |       |    |      |      |
| Leu        | UUA   | 785| 0.60 |      | GGC        | 123   | 0.19 |      |
|            | UUG   | 556| 0.42 | \textit{trnL-CAC} | GCC   | 586| 0.91 |      |
|             |       |    |      |      |             |       |    |      |      |
|             | CUA   | 371| 0.28 | \textit{trnL-UAG} | TYr   | 769| 1.59 |      |
|             | CUC   | 188| 0.14 |      | UAC        | 196   | 0.41 | \textit{trnY-GUA} |
|             | CUG   | 173| 0.13 |      | HIS        | 144   | 0.45 | \textit{trnH-GUG} |
|             | CUU   | 551| 0.42 |      | CAU        | 493   | 1.55 |      |
|             | AUA   | 718| 0.64 | \textit{trnL-CAG} | Gln   | 668| 1.49 | \textit{trnQ-UUG} |
|             | AUC   | 487| 0.43 | \textit{trnL-GAU} | GAG   | 226| 0.51 |      |
|             | AUU   | 1045| 0.93 |      | ASN        | 274   | 0.44 | \textit{trnN-GUU} |
|             | ATG   | 613| 1.00 | \textit{trn0IM-CAU} | AAA   | 988| 1.47 |      |
|             | AUG   | 487| 0.43 | \textit{trnL-GAU} | GAG   | 226| 0.51 |      |
|             | AGU   | 1045| 0.93 |      | ASN        | 274   | 0.44 | \textit{trnN-GUU} |
|             | AGC   | 104| 0.10 | \textit{trnS-GCU} | Glu   | 1009| 1.49 | \textit{trnE-UUC} |
|             | AGA   | 414| 0.40 |      | GAG        | 346   | 0.51 |      |
|             | UCA   | 440| 0.43 | \textit{trnS-UAG} | Cys   | 78    | 0.48 | \textit{trnC-GCA} |
|             | UCC   | 338| 0.33 | \textit{trnS-GGA} | UGU   | 245   | 1.52 |      |
|             | UCG   | 179| 0.17 | \textit{trnS-GGA} | UGU   | 245   | 1.52 |      |
|             | UCU   | 574| 0.56 |      | Arg        | 512   | 0.65 | \textit{trnR-UCU} |
|             | CCA   | 312| 0.59 | \textit{trnP-UGG} | AGG   | 161   | 0.20 |      |
|             | CCC   | 207| 0.39 |      | CGA        | 345   | 0.44 |      |
|             | CCG   | 131| 0.25 |      | CGC        | 89    | 0.11 |      |
|             | CUC   | 407| 0.77 |      | CGG        | 123   | 0.16 |      |
|             | Trr   | 407| 0.77 |      | CGG        | 123   | 0.16 |      |
|             | ACA   | 417| 0.64 | \textit{trnT-UAG} | CGU   | 344   | 0.44 | \textit{trnR-ACG} |
|             | ACC   | 241| 0.37 | \textit{trnT-GUA} | GLy   | 712   | 0.83 | \textit{trnG-UCG} |
|             | ACG   | 149| 0.23 |      | GCC        | 143   | 0.17 | \textit{trnG-GCC} |
|             | ACU   | 504| 0.77 |      | GGG        | 276   | 0.32 |      |
|             | GGU   | 587| 0.68 |      |            |       |    |      |      |

RSCU: Relative Synonymous Codon Usage.
doi:10.1371/journal.pone.0074736.t002
Table 3. Repeat sequences and their distribution in cpDNA of coconut.

| No. | Size (bp) | Start position | Repeat number | Type | Repeat sequence | Region |
|-----|-----------|---------------|---------------|------|-----------------|--------|
| 1   | 30        | 64504, 64537  | 2             | D    | TATACTATAATATATACTATAATAAATA | LSC; spacer between psbE and petL |
| 2   | 24        | 91629, 91653, 91677 | 3           | T    | GATATCGATATTGATATGATGAC  | IRB; ycf2 gene |
| 3   | 24        | 146981, 147005, 147029 | 3           | T    | ATATCGTACTATCATCATAATATCG | IRA; ycf2 gene |
| 4   | 21        | 149421, 149442 | 2             | T    | GAAGTGCATTGGAACAAAAAGA  | IRA; ycf2 gene |
| 5   | 20        | 31427, 31447  | 2             | T    | TTAAGAGATATACCTAGGAA  | LSC; spacer between trnT and psbD |
| 6   | 20        | 82734, 82754  | 2             | T    | CTGGTATACATTCATCAACAAAG | LSC; 3’ end of rps13 gene |
| 7   | 19        | 64518, 64537  | 2             | T    | TATACTATAATATATAATAAATA | LSC; spacer between psbE and petL |
| 8   | 17        | 12731, 12748  | 2             | T    | TCCTTTATTTTGTATTTG | LSC; intron of atpF gene |
| 9   | 13        | 28852, 28873  | 2             | D    | TATTATATATAAA  | LSC; spacer between petN and psbM |
| 10  | 13        | 59048         | 1             | I    | TATTATATAAA  | LSC; spacer between petN and psbM; spacer between accD and psal |
| 11  | 12        | 3749, 3773, 3793 | 3           | D    | AATATATATAATA | LSC; intron of trnK |
| 12  | 12        | 35106, 35118, 35141 | 3           | T    | ACTACTATACTA | LSC; spacer between trnG and trnFM |
| 13  | 12        | 35167         | 1             | I    | ACTACTATACTA | LSC; spacer between trnG and trnFM |

D: direct repeat; T: tandem repeat; I: inverted repeat.
*: inverted repeat sequence of repeat No. 9.
**: inverted sequence of repeat No. 12.

doi:10.1371/journal.pone.0074736.t003

Table 4. Comparison of repeat numbers and repeat lengths among 16 angiosperms.

| Taxon               | Total repeats (No.) | Longest repeat (bp) | References |
|---------------------|---------------------|---------------------|------------|
| **Monocots**        |                     |                     |            |
| Orchidaceae         |                     |                     |            |
| Cymbidium aloifolium| 232                 | 61                  | Yang et al. 2013 [29] |
| Arecales            |                     |                     |            |
| Cocos nucifera      | 13                  | 30                  | Produced in this study |
| Elaeis guineensis   | 7                   | 40                  | Uthaipaisanwong et al. 2012 [3] |
| Phoenix dactylifera | 11                  | 39                  | Yang et al. 2010 [2] |
| Poaceae             |                     |                     |            |
| Bamboo emeiensis    | 39                  | 132                 | Zhang et al. 2011 [39] |
| Hordeum vulgare     | 31                  | >55                 | Sasaki et al. 2007 [40] |
| Sorghum bicolor     | 26                  | >55                 | Sasaki et al. 2007 [40] |
| Agrostis stolonifera| 19                  | >55                 | Sasaki et al. 2007 [40] |
| **Dicots**          |                     |                     |            |
| Geraniaceae         |                     |                     |            |
| Geranium palmatum   | 100–150             | >200                | Guisinger et al. 2011 [4] |
| Pelargonium hortorum| ca. 200             | >200                | Guisinger et al. 2011 [4] |
| Rutaceae            |                     |                     |            |
| Citrus sinensis     | 29                  | 53                  | Bausher et al. 2006 [41] |
| Malvaceae           |                     |                     |            |
| Gossypium hirsutum  | 54                  | 72                  | Lee et al. 2006 [43] |
| Solanaceae          |                     |                     |            |
| Atropa belladonna   | 40                  | 45–49               | Daniell et al. 2006 [42] |
| Nicotiana tabacum   | 33                  | >55                 | Daniell et al. 2006 [42] |
| Solanum lycopersicum| 40                  | >55                 | Daniell et al. 2006 [42] |
| Solanum tuberosum   | 31                  | 50–54               | Daniell et al. 2006 [42] |

doi:10.1371/journal.pone.0074736.t004
eight rRNA genes while the latter is represented by pseudo genes contain introns; and four pairs of genes overlap (4 bp between 
\text{psbC} and \text{clpP}), which have two introns. Most protein-coding genes have an intron, except \text{psbB} and \text{clpP}, which have two introns. Most protein-coding genes have standard AUG as initiator codon; however, \text{psbD} and \text{ndhD} have an initiator codon of ACG, \text{rps19} starts with a GUG codon, and the initiator codon of \text{cemD} is ambiguous. The frequency of codon usage in the coconut cp genome is summarized in Table 2. Similar to many cp genomes of angiosperms [2,3,19–22], a strong bias toward an A or T in the third position of synonymous codons is observed in the coconut cp genome. The most and least prevalent amino acids are leucine (2624) and cysteine (323), respectively.

Although RT-PCR analysis validated that C-to-U editing changed the ACG start codon to AUG in the \text{ndhD} gene, the ACG start codon in the \text{rpl2} gene appeared to remain unedited in repeated experiments. However, we cannot eliminate the possibility that a low level of editing occurs in \text{rpl2}. Although less frequent than AUG, translation initiated at an ACG or GTG start codon is not unprecedented in plants. A previous study demonstrated that an initiator codon of ACG is not required to specify the initiation site for a proper translation in the cp genome [23]. GUG codons have been shown to be more efficient than AUG in initiating translation and have a relative strength varying from 15 to 30% of AUG activity [24]. In angiosperms, a GUG start codon has been found in the \text{cemD} gene [5,25–27] and \text{rps19} gene [2,3,5,8,26,28–32]. A transcript starting with an ACG start codon has been found in the \text{rps19} gene [5,25–27] and \text{cemA} gene [2,3,8,26,28–32]. A transcript starting with an ACG start codon is not unprecedented in plants. A previous study demonstrated that an initiator codon of ACG is not required to specify the initiation site for a proper translation in the cp genome [23]. GUG codons have been shown to be more efficient than AUG in initiating translation and have a relative strength varying from 15 to 30% of AUG activity [24]. In angiosperms, a GUG start codon has been found in the \text{cemD} gene [5,25–27] and \text{rps19} gene [2,3,5,8,26,28–32].

| Table 5. Comparison of cp genomes among six palm species. |
|---|---|---|---|---|---|
| Characteristics | Calamus | Pseudophoenix | Phoenix | Bismarckia | Elaeis | Cocos |
| Size (bp) | 157,270 | 157,829 | 158,462 | 158,211 | 156,973 | 154,731 |
| LSC | 85,525 | 85,736 | 86,198 | 86,390 | 85,192 | 84,230 |
| SSC | 17,595 | 17,587 | 17,712 | 17,459 | 17,639 | 17,391 |
| IR | 54,150 | 54,506 | 54,552 | 54,362 | 54,142 | 53,110 |
| GC content (%) | 37.36 | 37.32 | 37.23 | 37.47 | 37.40 | 37.44 |
| Total number of genes | 131 | 131 | 131 | 131 | 131 | 129 |
| Protein-coding genes | 85 | 85 | 85 | 85 | 85 | 84 |
| G+C (%) | 38 | 38 | 38 | 38 | 38 | 37 |
| rRNAs (bp) | 192,481 | 191,886 | 192,511 | 120,079 | 10,782 | 90,130 |
| G+C (%) | 53 | 53 | 53 | 53 | 53 | 53 |
| bases (bp) | 10,748 | 10,756 | 10,766 | 10,789 | 10,782 | 10,570 |
| Gene with intron(s) | 22 | 22 | 22 | 22 | 22 | 22 |
| Protein-coding genes | 14 | 14 | 14 | 14 | 14 | 14 |
| Number of Pseudogenes | 8 | 8 | 8 | 8 | 8 | 8 |

Repeats

With a criterion of 100% match in repeat copies, the tandem repeat finder identified 13 sets of repeats that are longer than 10 bp, including eight tandem repeats, three direct repeats, and two inverted repeats (Table 3). Three of the repeats are found in the \text{ycf2} genes, which are in the IR regions. The remaining repeats are similar to that found in date palm and oil palm. In fact, five of the repeats found in coconut (No. 2, 3, 6, 11 and 12 in Table 3) are shared by both oil palm and date palm, though the copy number may differ. In addition, repeats No. 5 and No. 8 in coconut are shared by both oil palm while repeats No. 4 and 13 are shared by date palm.

Repetitive sequences in cp genomes may recombine and induce rearrangements [35–37], which could play a crucial role in stabilization of cpDNA [38]. Compared with other angiosperms, cp genomes of the palm family generally have fewer and shorter repeats (Table 4). Of the 13 repeats found in coconut cpDNA, the longest is 30 bp. The oil palm cp genome has seven repeats and the longest is 40 bp [3] while date palm has 11 repeats and the longest is 39 bp [2]. By contrast, more than 20 repeats, with the longest extending up to 132 bp, were reported in Poaceae [39,40]. About 232 repeats, ranging from 30 to 61 bp in length, were reported in \text{Cymbidium} orchid [29]. In \text{Citrus}, 29 repeats with a range of 30 to 59 bp in length were detected [41]. In the \text{Solanaceae} family, as many as 42 repeats, with the most extensive being 56 bp, have been reported [42]. The cp genome of \text{Gossypium} has 54 repeats, with a longest one of 64 bp [43]. In the \text{Geraniaceae} family, some cp genomes contain up to 9% of repeats.
Table 6. RNA editing predicted by PREP-cp program and confirmed by RT-PCR.

| Gene | Nucleotide Position | Codon change | Editing position within codon | Amino acid change | PREP Predicted | RT-PCR results |
|------|---------------------|--------------|------------------------------|-------------------|---------------|----------------|
| accD | 154                 | CGG - TGG    | 1                           | R-W               | +             | -             |
|      | 794                 | TCG - TGG    | 2                           | S-L               | -             | +             |
|      | 1157                | TCA - TTA    | 2                           | S-L               | +             | +             |
|      | 1159                | CAT - TAT    | 1                           | H-Y               | +             | -             |
|      | 1403                | CCT - CTG    | 2                           | P-L               | +             | -             |
| atpA | 914                 | TCA - TTA    | 2                           | S-L               | +             | +             |
|      | 1148                | TCA - TTA    | 2                           | S-L               | +             | +/-           |
| atpB | 1184                | TCA - TTA    | 2                           | S-L               | +             | *             |
| atpF | 92                  | CCA - CTA    | 2                           | P-L               | +             | +/-           |
| atpI | 428                 | CCC - CTC    | 2                           | P-L               | +             | +             |
| ccsA | 647                 | ACT - ATT    | 2                           | T-I               | +             | -             |
| clpP | 82                  | CAT - TAT    | 1                           | H-Y               | +             | *             |
|      | 559                 | CAT - TAT    | 1                           | H-Y               | +             | *             |
| matK | 188                 | TCA - TTA    | 2                           | S-L               | +             | -             |
|      | 653                 | CCA - CTA    | 2                           | P-L               | +             | -             |
|      | 734                 | TTC - TTT    | 3                           | D-F               | -             | +             |
|      | 919                 | CAT - TAT    | 1                           | H-Y               | +             | -             |
|      | 1267                | CAC - TAC    | 1                           | H-Y               | +             | +             |
| ndhA | 50                  | TCG - TTT    | 2                           | S-L               | +             | +             |
|      | 476                 | TCA - TTA    | 2                           | S-L               | +             | +             |
|      | 566                 | TCA - TTA    | 2                           | S-L               | +             | +             |
|      | 961                 | CCT - CTT    | 1                           | P-S               | +             | +             |
|      | 1073                | TCC - TTC    | 2                           | S-F               | +             | -             |
| ndhB | 149                 | TCA - TTA    | 2                           | S-L               | +             | +/-           |
|      | 467                 | CCA - CTA    | 2                           | P-L               | +             | +             |
|      | 542                 | ACG - ATG    | 2                           | T-M               | +             | +             |
|      | 586                 | CAT - TAT    | 1                           | H-Y               | +             | +             |
|      | 704                 | TCC - TTC    | 2                           | S-F               | +             | +             |
|      | 737                 | CCA - CTA    | 2                           | P-L               | -             | +*            |
|      | 830                 | TCA - TTA    | 2                           | S-L               | +             | +             |
|      | 836                 | TCA - TTA    | 2                           | S-L               | +             | +             |
|      | 1112                | TCA - TTA    | 2                           | S-L               | +             | +             |
|      | 1193                | TCA - TTA    | 2                           | S-L               | +             | +             |
|      | 1255                | CAT - TAT    | 1                           | H-Y               | +             | +             |
|      | 1481                | CCA - CTA    | 2                           | P-L               | +             | +/-           |
| ndhD | 2                   | ACG - ATG    | 2                           | T-M               | +             | +/-           |
|      | 59                  | TCA - TTA    | 2                           | S-L               | +             | +             |
|      | 383                 | TCA - TTA    | 2                           | S-L               | +             | +             |
|      | 674                 | TCG - TTT    | 2                           | S-L               | +             | +             |
|      | 947                 | ACA - ATA    | 2                           | T-I               | +             | +             |
|      | 1193                | TCA - TTA    | 2                           | S-L               | +             | +             |
| ndhF | 1310                | TCA - TTA    | 2                           | S-L               | +             | +             |
|      | 62                  | TCA - TTA    | 2                           | S-L               | +             | +/-           |
|      | 290                 | TCA - TTA    | 2                           | S-L               | +             | +/-           |
|      | 392                 | TCC - TTC    | 2                           | S-F               | +             | +             |
|      | 442                 | CAT - TAT    | 1                           | H-Y               | +             | +             |
|      | 586                 | CTT - TTT    | 1                           | L-F               | -             | +             |
|      | 1393                | CAC - TAC    | 1                           | H-Y               | +             | -             |
| Gene       | Nucleotide Position | Codon change | Editing position within codon | Amino acid change | PREP Predicted | RT-PCR results |
|------------|---------------------|--------------|-------------------------------|------------------|----------------|----------------|
| ndhG       | 2093                | TCC - TTC    | 2                             | S-F              | +              | -              |
| ndhH       | 314                 | ACA - ATA    | 2                             | T-I              | +              | -              |
|            | 347                 | CCA - CTA    | 2                             | P-L              | -              | +              |
| ndhK       | 505                 | CAT - TAT    | 1                             | S-L              | +              | +              |
|            | 545                 | TCT - TTT    | 2                             | S-F              | -              | +/-*           |
|            | 726                 | TAC - TAT    | 3                             | Y-Y              | -              | -              |
| petB       | 131                 | TCG - TTG    | 2                             | S-L              | +              | +              |
|            | 372                 | GTC - GTT    | 3                             | S-L              | -              | -              |
|            | 518                 | ATG - ACG    | 2                             | M-T              | -              | -              |
|            | 677                 | TCA - TTA    | 2                             | S-L              | -              | -              |
| psaI       | 418                 | CGG - TGG    | 1                             | R-W              | +              | +*             |
|            | 611                 | CCA - CTA    | 2                             | P-L              | +              | +              |
|            | 80                  | TCT - TTT    | 2                             | S-F              | +              | +              |
|            | 85                  | CAT - TAT    | 1                             | H-Y              | +              | +              |
| rpl2       | 2                   | ACG - ATG    | 2                             | T-M              | +              | -              |
| rpl20      | 26                  | ACA - ATA    | 2                             | T-I              | +              | -              |
| rpl22      | 242                 | TCA - TTA    | 2                             | S-L              | +              | -              |
| rpl23      | 71                  | TCA - TTA    | 2                             | S-L              | -              | +/-*           |
|            | 89                  | TCT - TTT    | 2                             | S-F              | -              | +/-*           |
| rpoA       | 200                 | TCT - TTT    | 2                             | S-F              | -              | +              |
|            | 368                 | TCA - TTA    | 2                             | S-L              | +              | +              |
|            | 527                 | TCC - TTC    | 2                             | S-F              | +              | +              |
|            | 830                 | TCA - TTA    | 2                             | S-L              | +              | +              |
| rpoB       | 467                 | TCG - TTG    | 2                             | S-L              | +              | +/-            |
|            | 545                 | TCA - TTA    | 2                             | S-L              | +              | +              |
|            | 560                 | TCG - TTG    | 2                             | S-L              | +              | +              |
|            | 617                 | CCG - CTG    | 2                             | P-L              | +              | +/-            |
|            | 1994                | TCT - TTT    | 2                             | S-F              | +              | +              |
| rpoC1      | 41                  | CCA - CTA    | 2                             | P-L              | +              | +              |
|            | 511                 | CGG - TGG    | 1                             | R-W              | +              | +              |
|            | 617                 | TCA - TTA    | 2                             | S-L              | +              | +              |
|            | 1663                | CAT - TAT    | 1                             | H-Y              | +              | -              |
| rpoC2      | 1381                | CAT - TAT    | 1                             | H-Y              | +              | -              |
|            | 2275                | CGG - TGG    | 1                             | R-W              | +              | -              |
| rps2       | 2309                | TCG - TTG    | 2                             | S-L              | +              | +              |
|            | 134                 | ACA - ATA    | 2                             | T-I              | +              | +              |
| rps3       | 248                 | TCA - TTA    | 2                             | S-L              | +              | +              |
|            | 30                  | TTC - TTT    | 3                             | H-I              | -              | -              |
| rps7       | 470                 | ACA - ATA    | 2                             | S-L              | -              | +/-*           |
| rps8       | 583                 | CAT - TAT    | 1                             | S-L              | -              | +/-*           |
| rps14      | 149                 | CCA - CTA    | 2                             | P-L              | +              | +              |
|            | 30                  | TTC - TTT    | 3                             | H-I              | -              | -              |
higher) repetitive DNA [4,44] and many of the repeats are longer than 100 bp [4].

In view of the correlation between repetitive DNA content and sequence rearrangement, significant structural rearrangements are likely to be observed in cp genomes rich in repetitive sequences. This idea has been validated in many cases listed above such as Poaceae [35,39,40,42] and Geraniaceae [4,44–46]. Conversely, the relatively low content of repetitive DNA in cp genomes of the palm family suggests a relatively higher degree of stability and conservation across different palm species. Consistent with this notion, our investigation revealed neither significant recombination (Fig. S1) nor dramatic variation (Table 5) in the cp genomes of six palm species.

RNA editing sites

RNA editing is a posttranscriptional process that is mainly observed in mitochondrial and cp genomes of higher plants [47]. This process may induce the occurrence of substitution or indels, which in turn, can result in transcript alternation [33,47,48]. In coconut cpDNA, the PREP-cp program predicted 83 RNA editing sites out of 27 genes. Our RT-PCR analysis confirmed editing at 64 of those sites (Table 6). An additional six editing sites not predicted by the program were detected in accD, matK, ndhB, ndhG, ndhH, and rpoA. Of the genes investigated, ndh genes have the highest number of editing sites.

A comparative study of RNA editing across eight land plants demonstrated an evolutionary trend of decline (or complete loss) in the number of editing sites, silent editing, editing in the first or third position, and editing types other than C to U [47]. In angiosperms, the editing is almost exclusively a C to U substitution [49] and the total number of editing sites ranges from 20 to 37 [47,50–53]. Compared with other angiosperms, coconut has more than twice as many editing sites, although the editing characteristics are similar (Table 7). Moreover, because of the evolutionary conservation of RNA editing, closely related taxa usually share more editing sites [47]. For example, more editing sites are shared within Poaceae than those shared among grasses and dicots [54]. Similarly, related Nicotiana species share more editing sites with each other than with plants from other genera [34].

### Table 6. Cont.

| Gene | Nucleotide Position | Codon change | Editing position within codon | Amino acid change | PREP Predicted | RT-PCR results |
|------|---------------------|--------------|-------------------------------|-------------------|--------------|---------------|
| ycf1 | 3423 TAC → TAT | 3 Y-Y | - | - |
|      | 3429 GAT → GAC | 3 D-D | - | - |
|      | 3449 ATT → ACT | 2 I-T | - | - |
|      | 3852 ATC → ATT | 3 I-H | - | - |
|      | 4487 CTT → CCT | 2 L-P | - | - |
| ycf2 | 549 TCG → TCA | 3 S-S | - | - |
|      | 607 GAA → AAA | 1 E-K | - | - |
| ycf3 | 44 TCT → TTT | 2 S-F | + | + |
|      | 185 ACG → ATG | 2 T-M | + | + |
|      | 191 CCA → CTA | 2 P-L | + | + |
|      | 407 TCC → TTC | 2 S-F | + | + |
| ycf4 | 254 TCA → TTA | 2 S-L | - | +/-

*+/-*: editing; 
*+/+: partial editing; 
**+**/**: editing sites shared with oil palm [3].

doi:10.1371/journal.pone.0074736.t006

### Table 7. Comparison of RNA editing in six species of angiosperms.

| Arabidopsis | Nicotiana | Cocos | Elaeis | Zea | Oryza |
|-------------|-----------|-------|--------|-----|-------|
| Total editing sites | 34 | 37 | 32 | 26 | 21 |
| C to U editing (%) | 100 | 100 | 78.12 | 100 | 100 |
| U to C editing (%) | 0 | 0 | 0 | 0 | 0 |
| G to A editing (%) | 0 | 0 | 6.25 | 0 | 0 |
| Silent editing | 0 | 0 | 10 | 1 | 0 |
| Non-silent editing | 34 | 37 | 18 | 25 | 21 |
| Intron editing | 0 | 0 | 4 | 0 | 0 |
| 1st codon editing (%) | 14.7 | 5.4 | 16 | 15.4 | 4.8 |
| 2nd codon editing (%) | 85.3 | 91.9 | 82.67 | 46.1 | 92 |
| 3rd codon editing (%) | 0 | 2.7 | 1.33 | 23.5 | 4 |

doi:10.1371/journal.pone.0074736.t007
The rps19 pseudogenization and IR fluctuation

Dot plot analysis demonstrated that the gene content and organization of coconut cpDNA are nearly identical to other palm species (Fig. S1). Nevertheless, some variation could be detected. For instance, other palm species have two copies of the rps19 gene located near the IR A/LSC and IR B/SSC junctions respectively, whereas coconut has only one copy of rps19 at the IR B/SSC junction. At the IRA/LSC junction we found a rps19-like sequence of 174 bp, which is likely a pseudogene judged from its shorter length compared to the regular rps19 gene (279 bp). We speculate that the pseudogenization of the rps19 at IRA/LSC junction is due to IR fluctuation in coconut cpDNA.

A comparative study among cpDNAs of six palm species (Table 5) indicated that coconut has the smallest cp genome (154,731 bp) and the shortest IRs (53,110 bp). The largest cp genome with the longest IRs is found in Phoenix (158,462 bp and 54,552 bp, respectively). Similarly to other cp genomes [2,3], the palm cp genomes, including coconut, are all AT-rich. Graphical alignment showed that the IRs have both expanded and contracted during the evolution of the palm family, though dramatic changes were not detected (Fig. 2).

Fluctuations of the IR regions have occurred sporadically during the evolutionary history of angiosperms [35]. Two of the most extreme cases are found in Pelargonium hortorum of the Geraniaceae and a group of legumes that includes pea and broad beans. The single IR region has expanded to 76 kb [46] in the former whereas one copy of the IRs is completely lost from cp genomes of the latter [1]. The structurally conserved feature of the IR regions is resistant to recombinational loss [56]. The presence of the IR regions may thus help to stabilize the cp genome. The most direct evidence for this suggestion is that more rearrangements occurred within the group of legumes that have lost a copy of IR than those that have not [57]. Another piece of evidence is the acceleration of synonymous substitution rates in the remaining copy of the duplicated region [56]. Consequently, we can infer that the evolutionary rates of cp genomes in the palm family are relatively mild, judging from the comparatively minor fluctuation of the IR regions.

Phylogenetic analysis and events of gene gain and loss

Our phylogenetic reconstruction built upon 47 protein-coding genes of cp sequences, rooted by Amborella, supported three major monophyletic groups: magnoliids, monocots, and eudicots (Fig. 3). Within monocots, Acorus (Acorales) diverged from other monocots first, followed by Colocasia (Alismatales), then by Cymbidium (Asparagales), which is sister to a clade that forms a monophyletic group of commelinids. The commelinids contain two sister clades. Within the first clade, Arecales group with the family Dasypogonaceae. In the second clade, Poales is sister to a subclade, which includes Zingiberales and Commelinales (Fig. 3). This topology is
consistent with a phylogenetic study of commelinids based on 83 plastid genes [8]. Moreover, our inference of relationships within the Arecales is also congruent with a thorough study of the palm family using a supermatrix method with 16 data partition [58].

We then mapped the related gene duplication and pseudogenization events onto the tree according to parsimony criteria. Our results indicate that the duplication of \( \text{rps19} \) gene near the IR A/LSC junction likely occurred before the divergence of Asparagales from the remaining monocots, which consist of Arecales, a family (Dasypogonaceae) with indecisive order (Dasypogon and Kingia), Poales, Commelinales, and Zingiberales (Fig. 3). After the lineages differentiated, the duplicated \( \text{rps19} \) event eventually became a pseudogene independently in \( \text{Cocos} \) of the Arecales, \( \text{Heliconia} \) of the Zingiberales, and \( \text{Nandina} \) of the Ranunculales. It has been completely lost in \( \text{Xiphidium} \) of the Commelinales and \( \text{Ceratophyllum} \) of the Ceratophyllales (Fig. 3).

In monocots, the overlap between \( \text{ndhF} \) and \( \text{ycf1} \) was found in a clade that contains Arecales and Dasypogonaceae. However, it was also found in \( \text{Drimys} \) of the Canellales and \( \text{Chloranthus} \) of the Chloranthales, both belong to the magnoliids. Following the parsimony rule, we concluded that the occurrence of the overlap between \( \text{ndhF} \) and \( \text{ycf1} \) in monocots and magnoliids arose from three independent events.

In summary, we have presented here the first complete cp genome sequence from coconut palm. Although the cp genome of coconut is the smallest found so far among palms, it shares the same overall organization, gene content and repeat structure that have been observed with cpDNA sequenced from other palm species. Nevertheless, unique features were found for the coconut genome, including pseudogenization of \( \text{rps19} \)-like gene and an unusually high number of RNA editing sites. A closer relationship between coconut and oil palms than with date palm was supported by phylogenetic relationships among angiosperms. Our data will contribute to the growing number of molecular and genomic resources available for studying coconut palm biology.

**Supporting Information**

**Figure S1 Dot plot analysis.** The cp genomes are nearly identical in the palm family. (TIF)

**Table S1** Primers used for gap-filling PCR and RT-PCR. (DOCM)

**Acknowledgments**

We thank Mr. Chi-Tai Lin, a local dwarf coconut breeder, from Hengchun peninsula in southern Taiwan for providing coconuts.

**Author Contributions**

Conceived and designed the experiments: YYH AJMM MM. Analyzed the data: YYH. Contributed reagents/materials/analysis tools: YYH. Wrote the paper: YYH MM.

**References**

1. Palmer J (1985) Comparative organization of chloroplast genomes. Annu Rev Genet 19: 325–354.
2. Yang M, Zhang X, Liu G, Yin Y, Chen K, et al. (2010) The complete chloroplast genome sequence of date palm (Phoenix dactylifera L.). PloS ONE 5: e12762.
3. Uthaipaisanwong P, Chanprasert J, Shearnman JR, Sangtrakru D, Yoocha T, et al. (2012) Characterization of the chloroplast genome sequence of oil palm (Elaeis guineensis Jacq.). Gene 500: 172–180.
4. Guisinger MM, Kuehl JV, Boore JL, Jansen RK (2011) Extreme reconfiguration of plastid genomes in the angiosperm family Geraniaceae: rearrangements, repeats, and codon usage. Mol Biol Evol 28: 583–600.
32. Moore MJ, Dhingra A, Soltis PS, Shaw R, Farmerie WG, et al. (2006) Rapid evolutionary analysis of the chloroplast genome sequence of mangrove (Ungnaya radiata) determined by high-throughput Pyrosequencing: structural organization and phylogenetic relationships. DNA Res 17: 11–22.

31. Hansen DR, Dastidar SG, Cai Z, Penaflor C, Kuehl JV, et al. (2007) The complete chloroplast genome sequence of mangrove (Ungnaya radiata) determined by high-throughput Pyrosequencing: structural organization and phylogenetic relationships. DNA Res 17: 11–22.

30. Meerroor, A, Kruger R, Singh R, Low E-T, Ithim M, et al. (2012) Coconut, date fruit, and guava genetics. In: Schnell RJ, Priyadarsan KM, Genomics of Tree Crops: Springer New York. 299–311.

29. Yang JB, Tang M, Li HT, Zhang ZR, Li DZ (2013) Complete chloroplast genome sequence of mangrove (Ungnaya radiata) determined by high-throughput Pyrosequencing: structural organization and phylogenetic relationships. DNA Res 17: 11–22.

28. Raubeson L, Peery R, Chumley T, Dziubek C, Fourcade H, et al. (2007) Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. Proc Natl Acad Sci USA 104: 19369–19374.

27. Moore MJ, Bell CD, Soltis PS, Soltis DE (2007) Using plastid genome-scale data to resolve enigmatic relationships among angiosperms. Mol Phylogenet Evol 43: 815–834.

26. Raubeson L, Peery R, Chumley T, Dziubek C, Fourcade H, et al. (2007) Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. Proc Natl Acad Sci USA 104: 19369–19374.

25. Marchal A, Parent J, Veromme-Lafontaine F, Joycey A, Lang BF, et al. (2009) Whirly proteins maintain plastid genome stability in Arabidopsis. Proc Natl Acad Sci U S A 106: 14693–14698.

24. Zhang YJ, Ma P-F, Li D-Z (2011) High-throughput sequencing of six bamboo chloroplast genomes: phylogenetic implications for temperate woody bamboos (Poaceae: Bambusoideae). PLoS ONE 6: e20596.

23. Chen X, Kindle KL, Stern DB (1995) The initiation codon determines the location and direction of translation in mitochondrial genes. J Mol Biol 262: 241–249.

22. Delannoy E, Fujii S, Colas des Francs-Small C, Brundrett M, Small I (2011) The complete plastid genome sequence of fern (Pteridium aquilinum) determined by high-throughput Pyrosequencing: structural organization and phylogenetic relationships. DNA Res 17: 11–22.

21. Shimada H, Sugiura M (1991) Fine structural features of the chloroplast organelle of bottle gourd (Lagenaria siceraria). Arch Histol Cytol 54: 479–490.

20. Shimada H, Sugiura M (1991) Fine structural features of the chloroplast organelle of bottle gourd (Lagenaria siceraria). Arch Histol Cytol 54: 479–490.

19. Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, et al. (2000) Artemis: sequence visualization and annotation. Bioinformatics 16: 944–945.

18. Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 27: 573–580.

17. Rutherford K, Parkhill J, Crook J, Hornell T, Rice P, et al. (2000) Artemis: sequence visualization and annotation. Bioinformatics 16: 944–945.

16. Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 27: 573–580.

15. Moore MJ, Dhingra A, Soltis PS, Shaw R, Farmerie WG, et al. (2006) Improving the resolution of complex phylogenetic relationships using chloroplast genome sequences. BMC Evol Biol 6: 77.

14. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucl Acids Res 32: 1792–1797.

13. Chumley TW, Palmer JD, Mower JP, Fourcade HM, Calie PJ, et al. (2006) The chloroplast genome sequence of aroids (Araceae) determined by high-throughput Pyrosequencing: structural organization and phylogenetic relationships. J Mol Evol 63: 654–678.

12. Moore MJ, Dhingra A, Soltis PS, Shaw R, Farmerie WG, et al. (2006) Rapid and accurate pyrosequencing of angiosperm plastid genomes. BMC Plant Biol 6: 43.