

# SCIENTIFIC REPORTS



OPEN

## The mitochondrial genome map of *Nelumbo nucifera* reveals ancient evolutionary features

Songtao Gui<sup>1,\*</sup>, Zhihua Wu<sup>1,\*</sup>, Hongyuan Zhang<sup>1</sup>, Yinzhen Zheng<sup>1</sup>, Zhixuan Zhu<sup>1</sup>, Dequan Liang<sup>2</sup> & Yi Ding<sup>1</sup>

Received: 13 October 2015

Accepted: 28 June 2016

Published: 22 July 2016

*Nelumbo nucifera* is an evolutionary relic from the Late Cretaceous period. Sequencing the *N. nucifera* mitochondrial genome is important for elucidating the evolutionary characteristics of basal eudicots. Here, the *N. nucifera* mitochondrial genome was sequenced using single molecule real-time sequencing technology (SMRT), and the mitochondrial genome map was constructed after *de novo* assembly and annotation. The results showed that the 524,797-bp *N. nucifera* mitochondrial genome has a total of 63 genes, including 40 protein-coding genes, three rRNA genes and 20 tRNA genes. Fifteen collinear gene clusters were conserved across different plant species. Approximately 700 RNA editing sites in the protein-coding genes were identified. Positively selected genes were identified with selection pressure analysis. Nineteen chloroplast-derived fragments were identified, and seven tRNAs were derived from the chloroplast. These results suggest that the *N. nucifera* mitochondrial genome retains evolutionarily conserved characteristics, including ancient gene content and gene clusters, high levels of RNA editing, and low levels of chloroplast-derived fragment insertions. As the first publicly available basal eudicot mitochondrial genome, the *N. nucifera* mitochondrial genome facilitates further analysis of the characteristics of basal eudicots and provides clues of the evolutionary trajectory from basal angiosperms to advanced eudicots.

The size of the mitochondrial genome differs among angiosperm species, ranging from approximately 220 kb (*Brassica napus*)<sup>1</sup> to 11.3 Mb (*Silene conica*)<sup>2</sup>, reflecting the insertion of noncoding DNA<sup>3</sup>, including DNA of plastid and nuclear origin and DNA from horizontal gene transfer (HGT)<sup>4,5</sup>. Genomic organization is also variable across angiosperm evolution, reflecting frequent intramolecular or intermolecular homologous recombination mediated through repeats<sup>6</sup>. The organization, gene content and RNA editing of the mitochondrial genome differs over a significant scale across seed plants. This variation is useful for studying the evolution of both genome structures and sequences.

*Nelumbo nucifera* Gaertn. (Sacred lotus) is considered an evolutionary relic, which like *Ginkgo*, *Sequoia*, *Metasequoia* and *Liriodendron*, has survived since the Late Cretaceous period<sup>7</sup>. During the Early Cretaceous period, *N. nucifera* was a perennial aquatic plant that flourished during the middle Albian<sup>8,9</sup>. Currently, *N. nucifera* has been classified in the monotypic family Nelumbonaceae, which contains a single genus *Nelumbo*. This genus includes two species, *N. nucifera* and *N. lutea*. As a eudicot whose lineage emerged prior to the divergence of core eudicots<sup>10</sup>, *N. nucifera* provides new insights into the origin of eudicots. The nuclear<sup>11,12</sup> and chloroplast<sup>13</sup> genomes of *N. nucifera* have recently been released. However, no information on the *N. nucifera* mitochondrial genome has been reported. Thus, it is important to sequence the *N. nucifera* mitochondrial genome to reveal the evolutionary characteristics of this plant and provide clues concerning the evolutionary trajectory from basal angiosperms to advanced eudicots.

Third-generation sequencing through single molecule real-time sequencing technology (SMRT)<sup>14,15</sup> produces considerably longer (up to 30 kb) unbiased DNA sequences without PCR amplification<sup>16</sup>. This technology has previously been used in *de novo* assembly through the PacBio RS II platform<sup>17–21</sup>.

In the present study, using an optimized method for mitochondrial DNA isolation, we prepared *N. nucifera* mitochondrial DNA and sequenced the genome using SMRT technology. The mitochondrial genome map was constructed after *de novo* assembly and annotation of the sequence data. Our analyses provide insights into the evolution of gene content and order, RNA editing patterns, positively selected sites and chloroplast DNA insertions in core eudicots.

<sup>1</sup>State Key Laboratory of Hybrid Rice, Department of Genetics, College of Life Sciences, Wuhan University, Wuhan 430072, China. <sup>2</sup>Nextomics Biosciences Co., Ltd., Wuhan, 430075, China. \*These authors contributed equally to this work. Correspondence and requests for materials should be addressed to Y.D. (email: yiding@whu.edu.cn)

| Category            | Feature                     | Number | bp (%)          |
|---------------------|-----------------------------|--------|-----------------|
| genome (524,797 bp) | G + C                       |        | 252,718 (48.16) |
| Genes (total)       |                             |        |                 |
|                     | protein coding <sup>1</sup> | 43     | 35651 (6.8)     |
|                     | pseudogenes                 | 9      | 5541 (1.1)      |
|                     | rRNA <sup>2</sup>           | 6      | 11407 (2.2)     |
|                     | mt-derived tRNA             | 13     | 1054 (0.2)      |
|                     | cp-derived tRNA             | 7      | 506 (0.1)       |
|                     | ORF                         | 96     | 38062 (7.3)     |
| Introns             |                             |        |                 |
|                     | cis-spliced                 | 20     | 39780 (7.6)     |
|                     | trans-spliced               | 5      | n.a             |
| Repeats             |                             |        |                 |
|                     | Repeats(>500)               | 8      | 48898 (9.3)     |
|                     | Short Repeats(20<, <500)    | 191    | 16668 (3.2)     |
|                     | SSRs                        | 201    | 2628 (0.5)      |

**Table 1. The statistics of the features of the *Nelumbo nucifera* mitochondrial genome.** <sup>1</sup>Each of gene *ccmFN*, *rps19* and *sdh3* has two copies. <sup>2</sup>Each of the three ribosome RNAs (*rrn26*, *rrn18*, *rrn5*) has two copies.

## Results

***N. nucifera* mitochondrial DNA isolation and genome assembly.** Mitochondria were purified from *N. nucifera* etiolated seedlings after discontinuous sucrose gradient centrifugation and DNase I digestion. *Janus Green B* staining showed that most isolated mitochondria were intact (Supplementary Figure S1). The 260/230 and 260/280 ratios of isolated mtDNA were 2.08 and 1.93, respectively. Semi-quantitative PCR showed that the isolated DNA was pure enough to build a library for sequencing (Supplementary Figure S2).

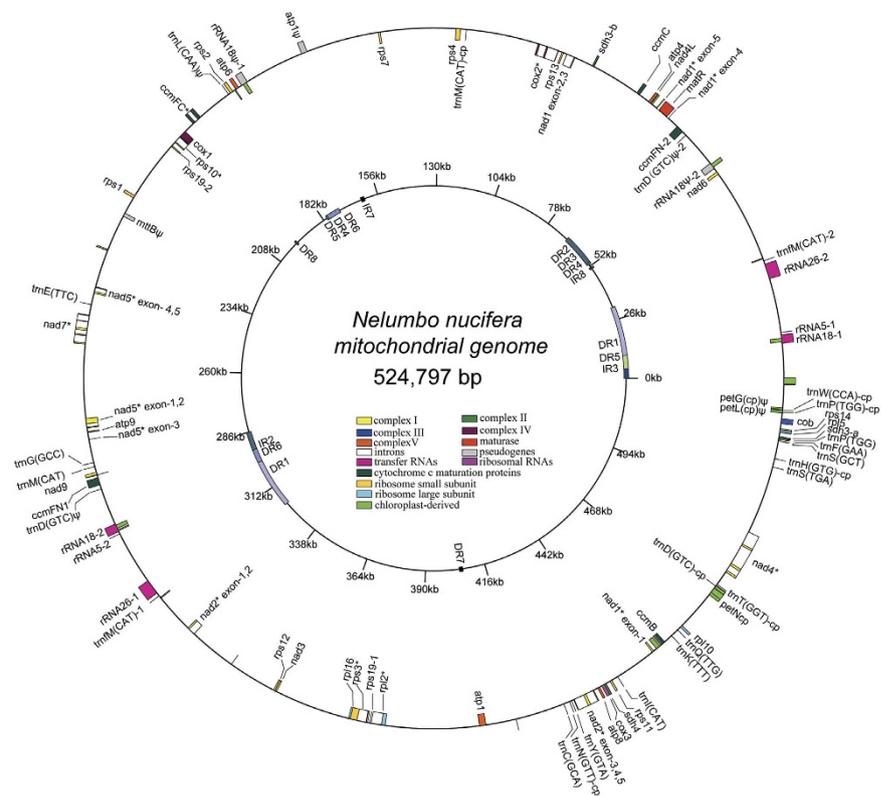
PacBio RSII sequencing generated 76,495 reads (341,866,338-bp in total), with a mean read quality of 0.83. After trimming off adapters and low quality regions and correcting by mapping short reads to long seeds, we have obtained 9,165 reads (42,623,117-bp in total, 4,651-bp per read on average) with an accuracy of 99%. After filtering chloroplast reads, a total of 7,151 reads (31,112,098-bp in total, 4,351-bp per read on average) were used for the assembly process, reaching a coverage depth of 59 over the *N. nucifera* mitochondrial genome. The assembly was verified by comparing with Sanger sequencing of PCR amplification using 18 primer pairs. ABI3730 sequencing generated a total of 20,176-bp sequences, representing 3.84% of the genome. Only one mismatch was detected at position 68,132 of the assembled mitochondrial genome (Supplementary Table S1), making the assembly accuracy of approximately 99.995%.

**Genome size and content.** The *N. nucifera* mitochondrial genome is assembled into a single circular-mapping<sup>22</sup> molecule of 524,797-bp (Table 1), with a GC content of 48.16%. To our knowledge, *N. nucifera* has the second highest GC content of all plant mitochondrial genomes, while the *Butomus umbellatus* mitochondrial genome has the highest GC content of 49.1%<sup>23</sup> (Supplementary Table S2). Eight long repeats (>500-bp) including four direct repeats (DRs) and four inverted repeats (IRs) were identified, accounting for 9.3% (48,898-bp) of the total size. In addition to the long repeats, the *N. nucifera* mitochondrial genome also contained many small repeats (20- to 500-bp), comprising 3.2% (16,668-bp) of the total length. Two hundred and one simple sequence repeats (SSRs) were identified (Supplementary Table S3), accounting for 0.5% (2628-bp) of the total length.

The *N. nucifera* mitochondrial genome contains a total of 63 genes, including 40 protein-coding genes, three rRNA genes (*rrn5*, *rrn18* and *rrn26*), 13 complete native mitochondrial tRNA genes and seven chloroplast-derived tRNA genes (Fig. 1, Table 2). Of these genes, *ccmFN*, *rps19* and all three rRNA genes have two identical copies, while *sdh3* has two different copies, *sdh3-a* and *sdh3-b*. *sdh3-a* is 39 bp longer than *sdh3-b* and has a 5-bp difference with *sdh3-b* at the 3'-terminal nucleotides (Supplementary Figure S3). Two different copies of a gene can result from horizontal gene transfer (HGT) events between angiosperm mitochondrial genomes<sup>5</sup> or mtDNA recombination. However, the high sequence similarity (100% identity from 1 to 331 bases) of *sdh3-a* and *sdh3-b* suggests that these two copies result from *N. nucifera* mtDNA recombination rather than HGT from other species. Ninety-six unknown functional open reading frames (ORFs) were also predicted in the present study, comprising 7.3% (38,062-bp) of the total length (Table 1). The *N. nucifera* mitochondrial genome contained 25 Group II introns, including 20 *cis*-spliced and five *trans*-spliced introns. The total length of the 20 *cis*-spliced introns is 39,780-bp, accounting for 7.6% of the entire *N. nucifera* mitochondrial genome (Table 1).

To analyse the potential codon bias in the *N. nucifera* mitochondrial genome, RSCU (relative synonymous codon usage)<sup>24</sup> was calculated using the coding sequences of 40 different mitochondrial protein-coding genes in *N. nucifera*. The results showed that A or U were more frequently used compared with G or C at the third position of *N. nucifera* mitochondrial codons (Supplementary Table S4), as observed in both chloroplast genomes and plant mitochondria genomes<sup>25,26</sup>. These results reflected a high A/T content at the third position of each codon<sup>27</sup>, that 61.27% of the *N. nucifera* mitochondrial codons have third position of A/T.

**Conserved gene clusters.** Based on the endosymbiont hypothesis that mitochondrial genomes are the remnants of a free-living prokaryote (most likely an extant  $\alpha$ -proteobacterium), some small DNA fractions of the



**Figure 1.** The *Nelumbo nucifera* mitochondrial genome map. Displayed as a circle, without implying that this is the *in vivo* conformation<sup>22</sup>. The genes are indicated in the outer circle, and the eight large repeats (>500-bp) with 99% identity are indicated in the inner circle. Chloroplast-derived regions are noted with a ‘-cp’ suffix. The same genes with different sequences are indicated with ‘-a,-b’. The same genes with the same sequences are indicated with ‘-1,-2’.

| Group of genes            | Name of genes  |
|---------------------------|--|
| Complex I                 | <i>nad1</i> * <sup>1</sup> , <i>nad2</i> *, <i>nad3</i> , <i>nad4</i> *, <i>nad4L</i> , <i>nad5</i> *, <i>nad6</i> , <i>nad7</i> *, <i>nad9</i>  |
| Complex II                | <i>sdh3-a</i> , <i>sdh3-b</i> , <i>sdh4</i>  |
| Complex III               | <i>cob</i>   |
| Complex IV                | <i>cox</i> , <i>cox2</i> , <i>cox3</i>   |
| Complex V                 | <i>atp1</i> , <i>atp4</i> , <i>atp6</i> , <i>atp8</i> , <i>atp9</i>  |
| Cytochrome c biogenesis   | <i>ccmB</i> , <i>ccmC</i> , <i>ccmFC</i> *, <i>ccmFN</i> (×2)  |
| Ribosome large subunit    | <i>rpl2</i> *, <i>rpl5</i> , <i>rpl10</i> , <i>rpl16</i>   |
| Ribosome small subunit    | <i>rps1</i> , <i>rps2</i> , <i>rps3</i> *, <i>rps4</i> , <i>rps7</i> , <i>rps10</i> *, <i>rps11</i> , <i>rps12</i> , <i>rps13</i> , <i>rps14</i> , <i>rps19</i> (×2)   |
| Intron maturase           | <i>matR</i>  |
| rRNA genes                | <i>rrn26</i> (×2), <i>rrn18</i> (×2), <i>rrn5</i> (×2)   |
| tRNA genes                | <i>trnC</i> (GCA), <i>trnE</i> (TTC), <i>trnG</i> (GCC), <i>trnK</i> (TTT), <i>trnM</i> (CAT), <i>trnFM</i> (CAT)(×2), <i>trnI</i> (CAT), <i>trnF</i> (GAA), <i>trnP</i> (TGG), <i>trnQ</i> (TTG), <i>trnS</i> (GCT), <i>trnS</i> (TGA), <i>trnY</i> (GTA) |
| chloroplast-derived genes | <i>petN</i> (cp), <i>trnM</i> (CAT)-cp, <i>trnN</i> (GTT)-cp, <i>trnD</i> (GTC)-cp, <i>trnI</i> (GGT)-cp, <i>trnH</i> (GTG)-cp, <i>trnP</i> (TGG)-cp, <i>trnW</i> (CCA)-cp   |
| Pseudogenes               | <i>atp1</i> ψ, <i>mttB</i> ψ, <i>petG</i> (cp)ψ, <i>petL</i> (cp)ψ, <i>ycf3</i> (cp)ψ, <i>rRNA18-1</i> ψ, <i>rRNA18-2</i> ψ, <i>trnD</i> (GTC)ψ, <i>trnL</i> (CAA)ψ  |

**Table 2.** List of the genes present in the mitochondrial genome of *Nelumbo nucifera*. <sup>1</sup>Genes with introns are indicated with asterisks (\*).

original genes have been identified in the mitochondrial DNA of higher order plants<sup>28</sup>. Therefore, conserved gene clusters remain present in some angiosperms despite multipartite organizations in angiosperm mitochondrial genomes<sup>29,30</sup>. A comparison of the collinearity among 38 analysed species revealed 15 collinear gene clusters in the *N. nucifera* mitochondrial genome relative to *Liriodendron tulipifera* (Supplementary Table S5): *rrn18-rrn5*, *atp4-nad4L*, *atp8-cox3-sdh4*, *rpl2-rps19-rps3-rpl16*, *nad1\_exon2,3-rps13*, *cox1-rps10*, *nad5\_exon4,5-trnE* (TTC)-*nad7*, *nad5\_exon1,2-atp9*, *nad3-rps12*, *trnN*(GTT)(cp)-*trnY*(GTA), *ccmB-trnK*(TTT)-*trnQ*(TTG), *trnS*(GCT)-*trnF*(GAA)-*trnP*(TGG)-*sdh3-a*, *cob-rps14-rpl5*, *trnP*(TGG)(cp)-*trnW*(CCA)(cp) and *petG*ψ-*petL*ψ.

Among these, six *N. nucifera* mitochondrial gene clusters were found conserved in other angiosperm species. *Atp8-cox3-sdh4* occurred in both the ‘early-diverging’ angiosperm *L. tulipifera* and some eudicot species. *Cox1-rps10*, *nad5\_exon1,2-atp9*, *trnN(GTT)(cp)-trnY(GTA)* and *petGψ-petLψ* were represented only in eudicots. *Nad5\_exon4,5-trnE(TTC)-nad7* was found only in *L. tulipifera* and *N. nucifera*. The nine remaining gene clusters were variously distributed among all of these taxa.

**RNA editing of *N. nucifera* mitochondrial protein-coding genes.** In the present study, with the exception of *rpl10*, all of the mitochondrial protein-coding genes of *N. nucifera* were successfully amplified, and a total of 33,719-bp cDNA sequences were sequenced; the amplicon for *ccmFC* was derived from pre-mRNA, with unspliced introns rather than mature mRNA. After aligning the DNA and cDNA sequences, 700 RNA editing sites were identified, including 612 nonsynonymous, 58 synonymous and 30 unpartitioned edits of the pseudogene *mttB* (Supplementary Table S6). As the same RNA editing sites were identified in both copies of *sdh3*, the RNA editing sites of the *sdh3* genes were only counted once. Approximately 87% of the total editing sites changed the translated amino acid sequence. Three U-to-C edits were identified in *ccmFN*, *cox1* and *cox2*, changing GUG (V) to GCG (A), UGC (C) to CGC (R) and UCU (S) to CCU (P), respectively. The most frequently edited genes were *atp9* and *ccmB*, with editing frequencies of 7.111 edits/100 nt and 6.924 edits/100 nt, respectively, while the frequencies of other genes were less than five edits/100 nt. The genes with no RNA editing sites included *ccmC*, *rps12*, *rps2* and *sdh4*. *N. nucifera* showed a pattern of relative editing levels across genes that was similar to other angiosperms<sup>31</sup>. For example, the ribosomal protein-coding genes had fewer edits, while the NADH dehydrogenase genes had more RNA edits.

In addition to RNA editing site analysis, we compared the RNA editing data obtained from *Amborella trichopoda*, *L. tulipifera*, *Vitis vinifera* and *Oryza sativa* with that of *N. nucifera* to identify similarities in the RNA editing sites (Supplementary Table S7). The results showed that the total number of RNA editing sites of *A. trichopoda*, *L. tulipifera* and *N. nucifera* (818, 802 and 700, respectively) were higher compared with *V. vinifera* and *O. sativa* (403 and 489, respectively). To examine similarities in RNA editing, we summarized the number of unique and shared edits in Supplementary Figure S4 using a pairwise comparison of the RNA editing sites of these five species. Compared with *A. trichopoda*, the maximum number of shared edits was 444 in *L. tulipifera*, followed by 361 in *N. nucifera*, 209 in *O. sativa* and 176 in *V. vinifera*. Compared with *L. tulipifera*, the maximum number of shared edits was 482 in *N. nucifera*, followed by 261 in *O. sativa* and 205 in *V. vinifera*. Moreover, *N. nucifera* contained the fewest number of unique edits (Supplementary Figure S4).

**Phylogenetic and selection pressure analysis.** To examine the phylogenetic evolution of *N. nucifera* mitochondria, a maximum likelihood (ML) tree of 79 plant mitochondrial genomes, including *N. nucifera*, was built using the sequences of 41 mitochondrial protein-coding genes and rooted with the green algae *Mesostigma viride*<sup>32</sup> (Supplementary Figure S5). The resulting tree showed that the topology was consistent with the representatives of the Angiosperm Phylogeny Group (APG) III<sup>33</sup>. In the phylogenetic tree, the algae were sisters of land plants; pteridophyte *Huperzia squarrosa* was a sister of the gymnosperm *Cycas taitungensis* and Angiospermae. Within the Angiospermae clade, *Amborella trichopoda* was sister of all other angiosperms. The monocots and the eudicots were monophyletic; *Spirodela polyrhiza* was sister of all other monocots; and *N. nucifera* formed an independent basal branch as a sister of all other eudicots. Each of these relationships has 100% bootstrap value support.

To explore potential horizontal gene transfer (HGT) events that occurred in *N. nucifera* mitochondria, Maximum Likelihood (ML) trees were reconstructed from 41 individual gene alignments using optimal substitution models (Supplementary Figure S6). The support of topological structure of each gene tree was evaluated by both bootstrap values of 1,000 bootstrap replicates and approximate likelihood ratio test (aLRT) values. The results showed that, among the 41 single gene trees, *N. nucifera* was located at the bottom of the angiosperm cluster as sister to the early-derived angiosperms *L. tulipifera* and *A. trichopoda* in 19 gene trees, and located as the basal branch of eudicot clusters in another 18 gene trees. In the rest four trees, *N. nucifera* was also embedded in the eudicot clusters: In *nad4L* and *rps19* trees, *N. nucifera* was sister to *V. vinifera*, which formed a basal branch of eudicot cluster in the mitochondria tree (Supplementary Figure S5). In *cox1* tree, *N. nucifera* was sister to *Vaccinium macrocarpon*, which was sister of all other Asterids in the mitochondria tree. And in *sdh4* tree, *N. nucifera* was sister to *Malus x domestica* which located at the bottom of Rosids cluster in the mitochondria tree. Many of the 41 gene trees had topologies inconsistent with the mitochondrial tree (Supplementary Figure S5) or the phylogenetic history of some plant species. For examples, *H. squarrosa* was embedded into the bryophytes cluster in 15 of all 33 gene trees containing *H. squarrosa*. And *Chara vulgaris* was clustered as sister to *Nitella hyalina* in all 32 gene trees containing *C. vulgaris*, 22 of which were well supported by both the bootstrap and aLRT values. Some eudicots were found among totally unrelated plant phyla in several gene trees: In *atp9* tree, *B. umbellatus* was clustered as sister to eudicots *Silene vulgaris* and *Silene latifolia* with bootstrap value of 85% and aLRT value of 0.89, but formed a long branch. In *nad7* tree, *V. macrocarpon* was embedded in the monocots cluster as sister to *Spirodela polyrhiza*. In *rps7* tree, *Salvia miltiorrhiza* and *Boea hygrometrica* were sisters to *Entransia fimbriata* with 100% bootstrap support and 100% aLRT support, but formed long branches. In *rps11* tree, *V. vinifera* was embedded into alga cluster. In *rpl2* tree, eudicots *S. miltiorrhiza*, *Ajuga reptans* and *Daucus carota* were clustered within alga branch. And in *sdh3* tree, eudicot *Milletia pinnata* was also embedded into alga cluster. Several observations may explain the cause of these tree abnormalities. The affinity of *S. miltiorrhiza* and *B. hygrometrica* with *E. fimbriata* in *rpl2* tree was more of a long branch attraction (LBA) of highly divergent sequences, because the alignment demonstrated that the *rps7* sequence of *E. fimbriata* was highly divergent from that of *S. miltiorrhiza* and *B. hygrometrica*. NCBI-BLASTN searches of these *rps7* genes against the Nucleotide Collection Database showed that *rps7* in *S. miltiorrhiza* and *B. hygrometrica* mitochondria have shown 100% identity with *rps7* of their own chloroplast genomes, while *rps7* in *E. fimbriata* showed low similarity with other nucleotides.

The abnormalities in *rps11* tree and *rpl2* tree could also be caused by genes of chloroplast origin, because BLASTN showed a high similarity of *V. vinifera* mitochondrial *rps11* with its chloroplast *rps11*, the *rpl2* of *S. miltiorrhiza*, *A. reptans* and *D. carota* also showed 100% similarities with their chloroplast ones. The *sdh3* of *M. pinnata* showed extremely low similarity with records in both nucleotide collection database and non-redundant protein database of NCBI, which leads to the embedding of *M. pinnata* into alga cluster with a extremely long branch. The abnormalities in *nad7* tree may be caused by lacking of informative sites due to the limited sequence variations, since the deletion of *S. polyrhiza* and *Phoenix dactylifera* leads to the relocation of *V. macrocarpon* as sister to *D. carota* in *nad7* tree with higher branch support values. And the alignments showed that *V. macrocarpon* shows higher similarity with *D. carota* than with *S. polyrhiza*. The abnormalities in *atp9* tree may be also caused by lacking of informative sites considering its short length (225-bp in the alignments) and the high similarity between eudicots and monocots.

Mitochondrial protein-coding genes are essential for the assembly of respiratory chain complexes and the translation of proteins. Therefore, these genes are subject to purifying selection pressures<sup>34</sup>. The improved branch-site model was used to identify the positively selected sites of 41 mitochondrial genes in *N. nucifera* using the alignments and gene trees of the 79 mitochondria genomes. After likelihood ratio tests (LRT), alternative models were significantly favoured in 27 genes. However, in other genes, alternative models were not preferable to the null model (Supplementary Table S8). Positively selected sites were identified in 14 genes after Bayes Empirical Bayes (BEB) analyses. Among these 14 genes, only two genes had positively selected sites that were significantly supported by posterior probability values: 96.1% and 95.6% of the 25th and 31st sites in *sdh4*, respectively, and 99.6% of 36th site in *cox1*. The low proportion of positively selected genes in *N. nucifera* suggested that most of the *N. nucifera* mitochondrial genes were conserved and remained stable during evolution. The positively selected sites in *cox1* and *sdh4* might imply functional variations. Thus, the structural analysis and functional verification of these two genes will be necessary in subsequent research.

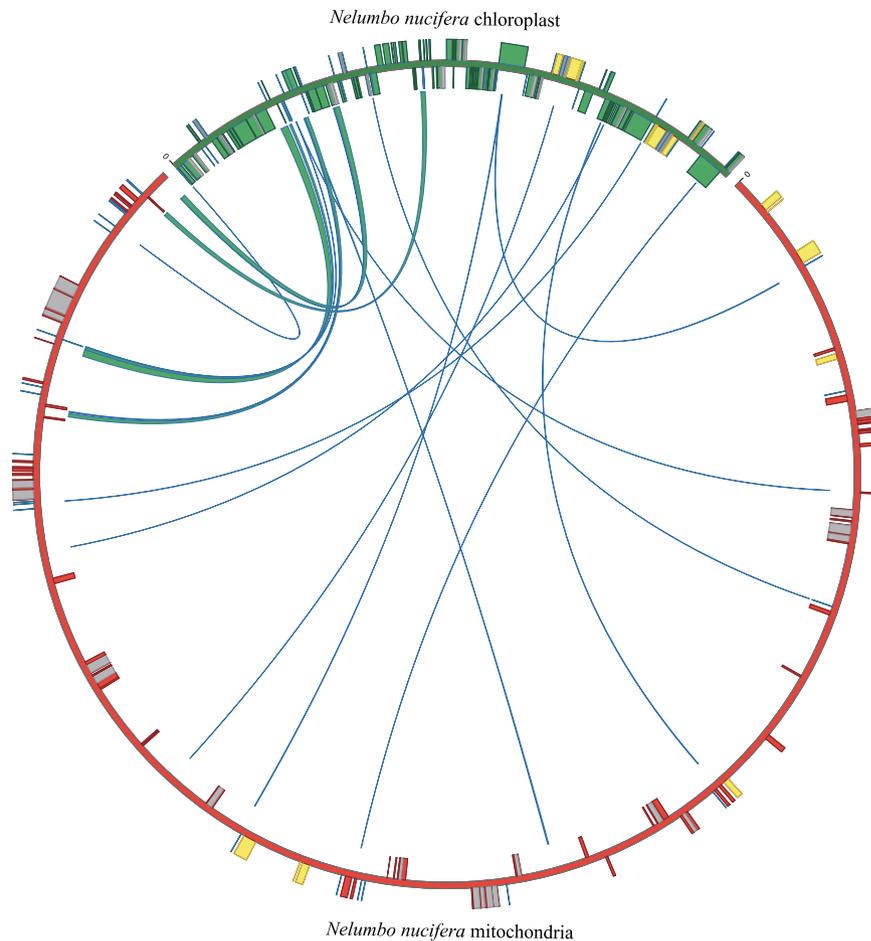
**Chloroplast DNA insertions in *N. nucifera* mtDNA.** Nineteen fragments of chloroplast DNA ranging from 54 to 1,998-bp were identified in *N. nucifera* mtDNA (Fig. 2, Table 3). The total size was 8,256-bp, approximately 1.6% of the *N. nucifera* mitochondrial genome and 5.0% of the chloroplast genome. After comparing the 19 chloroplast-derived fragments with the 78 other plant mitochondrial genomes used in the present study, homologues of 15 fragments were identified in other plant mtDNA (Supplementary Table S9). None of these chloroplast-derived fragments was observed in algae, bryophytes or the lycophyte *H. squarrosa*. The 141-bp fragment (positions 472035-472175) and the two shortest fragments (positions 91514-91569 and 342137-342190) were unique to the *N. nucifera* mitochondrial genome. The third shortest fragment (positions 413440-413496) was only identified in *A. trichopoda*. The longest fragment (positions 469237-471234) was only identified in *L. tulipifera*. The 84-, 81- and 78-bp fragments containing the *trnN*(GTT), *trnH*(GTG) and *trnM*(CAT) genes, respectively, had homologues in more than 40 species. The 84-bp fragment was commonly identified in angiosperms, but was absent in the gymnosperm *C. taitungensis*, while the 81- and 78-bp fragments were commonly observed in both angiosperms and *C. taitungensis*.

## Discussion

We sequenced the *N. nucifera* mitochondrial genome using third-generation SMRT sequencing technology and constructed a genomic map. In our previous work<sup>13</sup>, a comparison between *N. nucifera* chloroplast assemblies obtained using Sanger, Illumina MiSeq, and PacBio RS II platforms indicated that SMRT sequencing technology showed great promise for the production of highly accurate genome sequences because it produces long reads and lacks bias. In this study, the assembly validation experiment showed a great accuracy of *N. nucifera* mitochondrial genome assembly.

As the first publicly available basal eudicot mitochondrial genome, the *N. nucifera* mitochondrial genome provides valuable insights into the pattern of the evolution from basal angiosperms to advanced eudicots, and shows ancient evolutionary features in many aspects:

- (1) Conservation of genomic content. The *N. nucifera* mitochondrial genome retains 40 of all the 41 protein-coding genes present in ancestral flowering plant mitochondrial genomes<sup>35</sup>. This is consistent with previously described Southern blot results<sup>36</sup>, which suggest that the *Nelumbo* mitochondrial genome contains the full set of protein-coding genes present in the ancestral angiosperm. To our knowledge, of all available angiosperm mitochondrial genomes, only *A. trichopoda*<sup>5</sup> and *L. tulipifera*<sup>37</sup> have the full set of 41 protein-coding genes. Frequent loss and transfer of mitochondrial genes, particularly of the ribosomal protein-coding genes and succinate dehydrogenase (*sdh*) genes, has occurred in other angiosperm mitochondrial genomes and results in variable gene content among angiosperms<sup>36</sup>. The total length of the 20 *cis*-spliced introns of the *N. nucifera* mitochondrial genome is longer than any other sequenced angiosperm mitochondrial genome<sup>37</sup>. Currently, the complete set of 20 *cis*-spliced introns has only been identified in the mitochondria of three angiosperm species, including *N. nucifera* (in the present study), the 'early-diverging' angiosperm *L. tulipifera* and the 'early-diverging' core eudicot *Vitis vinifera*. In contrast, the five *trans*-spliced introns in *N. nucifera* were conserved in all of the sequenced angiosperm mitochondrial genomes.
- (2) Conservation of gene clusters. As a sister group, the 'early-diverging' angiosperm *L. tulipifera* occurred prior to the divergence of monocots and eudicots, showing primitive features in gene content and at RNA editing sites<sup>37</sup>. Hence, the *L. tulipifera* mitochondrial genome is a scale to explore the evolution of *N. nucifera* and other angiosperms. The *N. nucifera* mitochondrial genome has gene clusters that can date back to the original bacterial ancestor of mitochondria (*rrn18-rrn5*, *rpl2-rps19-rps3-rpl16*), gene clusters unique to angiosperms (*atp8-cox3-sdh4*, *nad1\_exon2,3-rps13*), a gene cluster unique to eudicots (*petGψ-petLψ*) and a gene cluster (*nad5\_exon4,5-trnE(TTC)-nad7*) that is only shared with *L. tulipifera*. This further indicates that the



**Figure 2. Schematic representation of chloroplast DNA transferred into the *Nelumbo nucifera* mitochondrial genome.** The green line within the circle shows the regions of the plastid genome that has been inserted into different locations of the mitochondrial genome. The genes in plastid and mitochondria are indicated by green and red boxes, respectively. The introns are indicated by grey boxes, and rRNAs and tRNAs are indicated by yellow and blue boxes, respectively. The boxes on the inside and outside of the maps were transcribed in clockwise and counterclockwise directions, respectively.

*N. nucifera* mitochondrial genome was more primitive than other advanced angiosperms in angiosperm evolution.

- (3) Retention of ancient RNA editing sites. During the evolution of advanced angiosperms, RNA editing sites showed various degrees of subsequent losses in different lineages<sup>37–39</sup>. Though the emergence and maintenance of RNA editing in plants has not yet been clearly explained, a hypothesis falling under the scheme of “constructive neutral evolution” holds that editing may have emerged through neutral processes, to replace functionally important cytosines that require post-transcriptional editing to produce a conserved amino acid<sup>40</sup>. According to this hypothesis, most editing sites are nonsynonymous. In *N. nucifera*, approximately 87% of the total editing sites changed the translated amino acid sequence. Comparing RNA editing sites of *A. trichopoda*, *L. tulipifera*, *V. vinifera* and *O. sativa* with that of *N. nucifera* clearly shows that the degree of editing loss in *N. nucifera* is much lower than that in *V. vinifera* and *O. sativa*, suggesting that the *N. nucifera* mitochondrial genome was more ancient.

*N. nucifera* was originally considered to be the closest relative to *Nymphaea alba* because of morphological similarities, reflecting convergent evolution with *N. alba* in the same aquatic environment<sup>41</sup>. However, phylogenetic analyses of angiosperms based on their nuclear genome have revealed that *N. nucifera* is a member of the Proteales within the basal eudicots and phylogenetic analyses using chloroplast sequences clustered *N. nucifera* into the basal eudicot branches as a sister to *Meliosma* and *Platanus*<sup>13</sup>. In this study, a phylogenetic tree built using the sequences of 41 mitochondrial protein-coding genes also puts *N. nucifera* as a sister to all other eudicots. Finally, the phylogenetic analysis in the present study, along with previous studies based on chloroplast and nuclear genome data, provided a comprehensive view of the basal location of *N. nucifera* in eudicots. Thus, we concluded that *N. nucifera* is an ‘early-diverging’ basal eudicot. In the most of the 41 gene trees, the *N. nucifera* sequences were clustered in the basal location in eudicots, suggesting that *N. nucifera* mitochondrial genes maintained low evolutionary rates. Topological incongruence was observed in different species of many single gene

| No. | Length | Position      | Genes Contained                      |
|-----|--------|---------------|--------------------------------------|
| 1   | 1998   | 469237–471234 | <b>petN;psbM</b>                     |
| 2   | 1387   | 523411–524797 | <i>ycf3</i>                          |
| 3   | 1229   | 516533–517761 | <i>petL-petG-trnW(CCA)-trnP(UGG)</i> |
| 4   | 997    | 451200–452196 | <i>psbC</i>                          |
| 5   | 621    | 471275–471895 | <b>trnD(GTC)</b>                     |
| 6   | 321    | 230963–231283 | <i>psbC</i>                          |
| 7   | 297    | 452532–452828 | <i>psbD;psbC</i>                     |
| 8   | 211    | 181488–181698 | <i>ndhD</i>                          |
| 9   | 206    | 29058–29263   | <i>ycf2</i>                          |
| 10  | 206    | 318795–319000 | <i>ycf2</i>                          |
| 11  | 145    | 523258–523402 | <i>ycf3</i>                          |
| 12  | 141    | 472035–472175 | <b>trnT(GGT)</b>                     |
| 13  | 87     | 285719–285805 | <i>ycf2</i>                          |
| 14  | 84     | 426801–426884 | <b>trnN(GTT)</b>                     |
| 15  | 81     | 505480–505560 | <b>trnH(GTG)</b>                     |
| 16  | 78     | 123420–123497 | <b>trnM(CAT)</b>                     |
| 17  | 57     | 413440–413496 | <i>ndhE</i>                          |
| 18  | 56     | 91514–91569   | <i>trnT(GGT)</i>                     |
| 19  | 54     | 342137–342190 | <i>trnA(TGC)</i>                     |

**Table 3. Chloroplast Insertions in the *Nelumbo nucifera* mitochondrial genome.** Pseudogenes were shown in regular font. Genes with intact ORFs and genes which encode tRNAs with intact folding ability are highlighted in bold.

trees in the present study. The embedding of *H. squarrosa* into the bryophytes clusters confirmed Liu's study on *H. squarrosa* mitochondria<sup>42</sup> which suggests the mitochondrial genome as the most archaic form in vascular plants and shows a gene content resembling those of charophytes and most bryophytes. *C. vulgaris* is considered to be the most closely related alga to land plants<sup>43</sup>. Turmel's study on *C. vulgaris* mitochondria<sup>43</sup> shows that *C. vulgaris* mtDNA strikingly resembles *Marchantia polymorpha* mtDNA, supported by both genome comparisons and phylogenetic analyses. The sister-group relationship between *C. vulgaris* and *N. hyalina* in our study suggested that *C. vulgaris* mtDNA may have more similarities with *N. hyalina* mtDNA. The embedding of species into totally unrelated plant phyla identified in six gene trees was expected to be possible HGT candidates. But our further analysis showed that this topological incongruence was caused either by long branch attraction of highly divergent genes, or by lacking of informative sites due to the limited sequence variations. Of the eight highly divergent genes, six genes were found similar to their own chloroplast genes, and three of these six genes were reported as chloroplast-derived by former studies (*rps11* of *V. vinifera*<sup>6</sup>, *rpl2* of *A. reptans*<sup>44</sup> and *rpl2* of *D. carota*<sup>4</sup>). If assuming correct genome assembly and annotation, the low similarity of *sdh3* of *M. pinnata* mitochondria compared with other nucleotide in the NCBI nucleotide collection database may show some new insights. However, in the present study, we found none of the topological incongruences showed a reliable evidence of HGT event. And among the 41 gene trees, *N. nucifera* was clustered with either eudicots or the early-derived angiosperms, indicating that HGT from other plant mitochondria rarely occurred in *N. nucifera* mitochondrial genes.

However, intracellular gene transfer (IGT) of DNA from the plastid to the mitochondrial genomes of higher plants is a common phenomenon on an evolutionary timescale. Sometimes, these IGT events may initiate a gain of functional tRNAs. Due to the slow rate of mitochondrial sequence change, some of the chloroplast-derived fragments identified in *N. nucifera* mitochondria may trace back to retention of an earlier HGT event. The absence of the 19 chloroplast-derived fragments in algae, bryophytes and the lycophyte *H. squarrosa* indicates that DNA transfer from the chloroplast genome to the mitochondrial genome in *N. nucifera* might have occurred after the gymnosperm/angiosperm divergence. We found that the *trnN(GTT)*-contained chloroplast-derived fragment was absent in the gymnosperm *C. taitungensis* while the *trnH(GTG)*- and *trnM(CAT)*-containing fragments were commonly observed in both angiosperms and *C. taitungensis*. This finding is consistent with the analysis of chloroplast-derived tRNA gene evolution<sup>37</sup>. Thus, it is likely that the chloroplast-derived genes *trnH(GTG)* and *trnM(CAT)*, previously present in *C. taitungensis*, were obtained from an earlier transfer, while *trnN(GTT)* might have occurred after the separation of gymnosperms and angiosperms. The short stretches of these three fragments in *N. nucifera* suggested that the selection of the flanking sequences might have occurred prior to the separation of early-diverging angiosperms and basal eudicots. Plant mitochondrial tRNAs were derived from three origins: (i) 'native' mitochondrial tRNAs, orthologous to those in moss mitochondria and derived from a mitochondrial ancestor; (ii) chloroplast-like mitochondrial tRNAs, highly homologous to chloroplast DNA; and (iii) unknown-origin mitochondrial tRNAs, such as the bacterial-like *trnC(GCA)*, obtained through horizontal gene transfer events<sup>45,46</sup>. Compared to all sequenced angiosperm mitochondrial genomes, the *N. nucifera* mitochondrial genome had a complete set of 13 intact 'native' mitochondrial tRNAs, which were also identified in the 'early-diverging' core eudicot *V. vinifera*, but variously presented in other angiosperms. The 'native' mitochondrial tRNAs retained in gymnosperms, *trnL(CAA)* and *trnD(GTC)*, have become pseudogenes in *N. nucifera* during the evolution of angiosperms<sup>47</sup>. The absence of *trnL(CAA)* and *trnD(GTC)* was

compensated with chloroplast-derived *trnL*(CAA) or cytosolic tRNAs<sup>3</sup>. The *N. nucifera* mitochondrial genome contained seven chloroplast-derived tRNA genes present in the ‘early-diverging’ core eudicot *V. vinifera*. Notably, the chloroplast-derived *trnT*(GGT) gene included in the 141-bp unique fragment in *N. nucifera* was also identified within a 3,571-bp chloroplast-derived fragment in *V. vinifera* mitochondria<sup>6</sup> but was absent in other plant mitochondrial genomes, suggesting that the gain of the chloroplast-derived *trnT*(GGT) gene in *N. nucifera* and *V. vinifera* mitochondria were separate events.

Though dozens of flowering plant mitochondrial genomes are published, early-derived angiosperms or basal eudicot mitochondrial genomes are still limited. The difficulty of isolating mitochondrial DNA is a bottleneck in acquiring complete mitochondrial genome sequence data. The optimized method for the isolation of mitochondrial DNA from plants rich in secondary metabolites used in the present study would promote the sequencing of more plant mitochondria. The genomic data of other basal eudicots, such as *Platanus*, will provide more detailed insights into mitochondrial genome evolution in basal eudicots.

## Materials and Methods

**Mitochondrial genome DNA isolation.** Mature seeds of *N. nucifera* were harvested from the Wuhan Moshan Botanical Garden (114°22′ N, 30°33′ E) of Hubei Province, China and germinated in the dark at room temperature. Discontinuous sucrose gradient centrifugation combined with DNase I digestion *in vitro* was used to separate mitochondrial DNA from etiolated seedlings. Three pairs of primers specific to genes in the nuclear, mitochondria and chloroplast were designed to evaluate the purity of the mtDNA through semi-quantitative PCR. A full description of the mtDNA isolation method has been provided in Supplementary Method S1.

**Library construction and sequencing.** A total of 20 µg of standard mtDNA was required to construct a size-selected ~20 kb SMRT-bell library. Subsequently, the SMRT-bell library was sequenced using three SMRT cells (Pacific Biosciences) with C2 chemistry on a PacBio RS II sequencing platform<sup>14</sup>. The sequencing and *de novo* assembly were performed at Yale University, Connecticut, America and Nextomics, Wuhan, China, respectively.

**Genome assembly and annotation.** The primitive reads from the Pacbio RS II were filtered and corrected by mapping short reads to long “seed” reads (6,000-bp) using SMRT Analysis 2.1 (Filtering parameters: Minimum Subread Length = 500, Minimum Polymerase Read Quality = 0.8, and Minimum Polymerase Read Length = 100. Correcting parameters: Alignment Candidates Per Chunk = 10, Total Alignment Candidates = 24, Number Of Seed Read Chunks = 6, Blasr: -noSplitSubreads -minReadLength 200 -maxScore -1000 -maxLCPLength 16 ). Then, the reads were mapped to both the *Nelumbo nucifera* chloroplast genome [GenBank:KM655836] and the NCBI chloroplast genome database using Blat<sup>48</sup> with default settings to filter chloroplast reads, reads with more than 90% matched and identity of more than 90% were removed. Both Celera Assembler 7.0<sup>49</sup> and MIRA 4.0rc5 were used for the assembly process. Reads longer than 8,000-bp were extracted for assembly. In the Celera Assembler assembly process, the fastqToCA utility in Celera Assembler was used to generate wrapper LIB messages, reads were assembled into contigs using runCA (-p asm -d asm -s asm.spec) and then aligned to the contigs using the BWA-SW algorithm in BWA-0.7.3a<sup>50</sup> with a mismatch penalty set to 5, contigs with less than 10X coverage depth were filtered. In the MIRA assembly process, the assembly was performed with parameters of “-GE:not = 20 -hirep\_best -SK:mchr = 30 -HS:mnr = yes -HS:nrr = 5 PCBIOHQ\_SETTINGS -CL:pec = no technology = PCBIOHQ”, redundancy was subsequently removed using self-to-self alignment performed by LastZ<sup>51</sup>. Both assembly processes output four contigs, these contigs were subsequently assembled into one contig using LastZ. At last, Quiver was used to correct the error regions and generate the best consensus sequence for the mitochondrial genome<sup>16</sup>. Default parameters were otherwise employed for processes above. To verify the assembly, 18 primer pairs were designed to target random selected intergenic regions (Supplementary Table S1).

The *N. nucifera* mitochondrial genome was manually annotated based on the output of NCBI-BLASTN and NCBI-BLASTX searches to a custom local database. The online tool tRNAscan-SE<sup>52</sup> was also used to search the tRNA genes. Conserved ORFs were identified through BLASTX searches of the NCBI database. The mitochondrial genome map of *N. nucifera* was drawn using Genome Vx<sup>53</sup>. Large repeats (>500-bp) were identified through BLASTN searches with 99% identities. Chloroplast insertions were identified through BLASTN searches of the *N. nucifera* chloroplast genome with 70% identity and an e-value of 1e-5. To identify the nuclear-derived sequences, the protein-coding genes, cis-spliced introns, rDNAs, tRNAs and chloroplast-derived sequences were masked; subsequently, the masked sequences were searched in the *N. nucifera* nuclear genome. The RSCU (relative synonymous codon usage) of the *N. nucifera* mitochondrial genome was calculated from the coding sequences of 40 different protein-coding genes. MISA<sup>54</sup> was used to screen for perfect SSRs, including mono-, di-, tri-, tetra-, penta- and hexa-nucleotide repeats. The minimum number of repeats was set at 10, 5, 4, 3, 3 and 3, respectively.

**Mitochondrial genome gene collinear analysis.** To evaluate the collinear relationships between genes, all protein-coding genes, tRNA genes and rRNA genes from the *N. nucifera* mitochondria genome were compared to 37 representative species of angiosperm mitochondria genomes available in NCBI using BLASTP/BLASTX<sup>55</sup>. The blast results were inputted into MCscan<sup>56</sup> with default parameters to compute multiple synteny. The final gene collinearity results were generated from the MCscan output file by swapping the gene order number of each gene with their names using a Perl script.

**RNA editing of *N. nucifera* mitochondrial protein-coding genes.** Total RNA was isolated from tender leaves using the modified CTAB method<sup>57</sup> (Supplementary Figure S7, a); total cDNA was produced through RT-PCR (reverse transcription PCR) (Supplementary Figure S7, b). The genome sequence data were used to

design PCR primers for all 40 protein-coding genes and one pseudogene (Supplementary Table S10). These primers were immediately flanked by genes based on the assumption that these regions comprised parts of the 5' and 3' untranslated regions, and the predicted PCR products were searched in the expressed sequence tags database using NCBI-BLASTN to evaluate whether the primers were in the untranslated regions. These genes were amplified from both the total DNA and total cDNA using the following conditions: 35 cycles of (40 s at 94 °C, 40 s at 48–56 °C and 1.5–2 min at 72 °C), with an initial step of 5 min at 94 °C and a final step of 10 min at 72 °C. The amplicons were detected on 1% agarose gels and purified using a Gel Extraction Kit (BioDev-Tech, Beijing, China). The purified cDNA amplicons were directly sequenced using an ABI 3730, and some of the purified products were also ligated into the pMD18-T Easy vector (TaKaRa, Dalian, China), transformed into Top10 chemically competent *E. coli* and subsequently sequenced. The sequenced cDNAs were compared with the genomic sequence and aligned using ClustalW<sup>58</sup> implemented in MEGA5<sup>24</sup>. The editing sites were detected according to the alignments and the chromatogram peaks of the cDNA sequences and subsequently partitioned based on a nonsynonymous or synonymous effect using a Perl script. The RNA editing data of *A. trichopoda*<sup>5</sup>, *L. tulipifera*<sup>37</sup>, *O. sativa*<sup>59</sup> and *V. vinifera*<sup>6</sup> were acquired from GenBank or REDIdb and parsed using Perl scripts.

**Phylogenetic and selection pressure analysis.** In addition to the sequenced mtDNA of *N. nucifera*, the protein-coding sequences from 78 published complete mitochondrial sequences of angiosperms were downloaded from the NCBI database (Supplementary Table S11). All 41 protein-coding genes were selected for phylogenetic analysis. These nucleotide sequences were separately aligned using ClustalW and manually adjusted, and the independent alignments were subsequently concatenated. The program jModelTest<sup>60</sup> was used to identify the most appropriate substitution model and combination of gamma distribution and proportion of invariant sites, with the parameter number of rate categories set to 4. Both the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) were calculated. During the analysis, if the predicted models of AIC and BIC were not the same, AIC models were used (instead of BIC models). Phylogenetic analyses of each gene and the concatenated sequences were performed using the maximum likelihood (ML) method implemented in RAxML-HPC v8.1.2<sup>61</sup> according to the optimal substitution models under the rapid bootstrap algorithm (1000 replicates). To test support for the branch points of each gene trees, nonparametric branch support tests based on the Shimodaira-Hasegawa-like approximate likelihood ratio test (SH-like aLRT) procedure were performed using phyML<sup>62</sup> with the tree topology search operation option of Nearest Neighbour Interchange (NNI), and the same substitution models as used in RAxML trees. R packages ape<sup>63</sup> and ggtree<sup>64</sup> were used to combine the branch supports of bootstrap value with SH-like aLRT value in each gene tree and present the tree structures. The alignments can be found in Supplementary Data S1, the results of jModelTest and all the trees can be found in Supplementary Data S2.

The program Codeml in the PAML package<sup>65</sup> was used to identify the positively selected genes in the *N. nucifera* mitochondrial genome. The phylogenetic trees of 41 protein-coding genes were used as tree structure files in Codeml, and *N. nucifera* was indicated as foreground in all gene trees. Test 2 of Model A in the branch-site model was selected, using the following parameters: in the null hypothesis, model = 2, NSSites = 2, fix\_omega = 1, omega = 1; in the alternative hypothesis, model = 2, NSSites = 2, fix\_omega = 0, omega = 1.5; and other parameters were left by default. Likelihood ratio tests (LRT) were performed, and the P value was calculated using the chi2 program in PAML packages.

**Chloroplast insertions in *N. nucifera* mtDNA.** After sequence annotation, the protein-coding genes, tRNA genes and rRNA genes of the *N. nucifera* mitochondrial genome were masked to eliminate BLAST hits of the conserved household genes in chloroplast and mitochondrial genomes. The masked genome sequence was BLASTed against the chloroplast genome of *N. nucifera*. The chloroplast-derived fragments were subsequently BLASTed against a local database containing all Viridiplantae mitochondrial genomes available in NCBI to identify homologues with an identity greater than 80% and full-length or nearly full-length (more than 90%) alignment. The maps of *N. nucifera* mitochondrial and chloroplast genomes and the chloroplast fragment transferences were drawn using Circos v.0.67<sup>66</sup>.

**Availability of supporting data.** The precise mitochondrial genome of *N. nucifera* has been submitted to GenBank [GenBank: KR610474]. The raw sequence data have been deposited in the Short Read Archive (SRA) database of NCBI [SRA: SRP066826], note that only data from SMRT cell 3 (named “mt\_s3.fastq”) was used for the assembly in this study. Other data sets supporting the results of the present study are included within the article and supplementary information.

## References

- Handa, H. The complete nucleotide sequence and RNA editing content of the mitochondrial genome of rapeseed (*Brassica napus* L.): comparative analysis of the mitochondrial genomes of rapeseed and Arabidopsis thaliana. *Nucleic Acids Res* **31**, 5907–5916 (2003).
- Sloan, D. B. *et al.* Rapid evolution of enormous, multichromosomal genomes in flowering plant mitochondria with exceptionally high mutation rates. *PLoS Biol* **10**, e1001241 (2012).
- Kubo, T. & Newton, K. J. Angiosperm mitochondrial genomes and mutations. *Mitochondrion* **8**, 5–14 (2008).
- Iorizzo, M. *et al.* De novo assembly of the carrot mitochondrial genome using next generation sequencing of whole genomic DNA provides first evidence of DNA transfer into an angiosperm plastid genome. *BMC Plant Biol* **12**, doi: 10.1186/1471-2229-12-61 (2012).
- Rice, D. W. *et al.* Horizontal transfer of entire genomes via mitochondrial fusion in the angiosperm Amborella. *Science* **342**, 1468–1473 (2013).
- Goremykin, V. V., Salamini, F., Velasco, R. & Viola, R. Mitochondrial DNA of *Vitis vinifera* and the issue of rampant horizontal gene transfer. *Mol Biol Evol* **26**, 99–110 (2009).
- Hsu, J. Late Cretaceous and Cenozoic vegetation in China, emphasizing their connections with North America. *Ann Mo Bot Gard* **70**, 490–508 (1983).

8. Upchurch, Jr. G. R., Crane, P. R. & Drinnan, A. N. The megaflores from the Quantico locality (Upper Albian), lower cretaceous Potomac group of Virginia. *Mem Virginia Mus Nat Hist* **4**, 1–57 (1994).
9. Crane, P. R. & Herendeen, P. S. Cretaceous floras containing angiosperm flowers and fruits from eastern North America. *Rev Palaeobot Palynol* **90**, 319–337 (1996).
10. Moore, M. J., Soltis, P. S., Bell, C. D., Burrell, J. G. & Soltis, D. E. Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots. *Proc Natl Acad Sci USA* **107**, 4623–4628 (2010).
11. Ming, R. *et al.* Genome of the long-living sacred lotus (*Nelumbo nucifera* Gaertn.). *Genome Biol* **14**(5), 1–11 (2013).
12. Wang, Y. *et al.* The sacred lotus genome provides insights into the evolution of flowering plants. *Plant J* **76**, 557–567 (2013).
13. Wu, Z. *et al.* A precise chloroplast genome of *Nelumbo nucifera* (Nelumbonaceae) evaluated with Sanger, Illumina MiSeq, and PacBio RS II sequencing platforms: insight into the plastid evolution of basal eudicots. *BMC Plant Biol* **14**, doi: 10.1186/s12870-014-0289-0 (2014).
14. Quail, M. A. *et al.* A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* **13**, doi: 10.1186/1471-2164-13-341 (2012).
15. Roberts, R. J., Carneiro, M. O. & Schatz, M. C. The advantages of SMRT sequencing. *Genome Biol* **14**, doi: 10.1186/gb-2013-14-7-405 (2013).
16. Chin, C. S. *et al.* Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods* **10**, 563–569 (2013).
17. Pugh, T. J. *et al.* Medulloblastoma exome sequencing uncovers subtype-specific somatic mutations. *Nature* **488**, 106–110 (2012).
18. Ferrarini, M. *et al.* An evaluation of the PacBio RS platform for sequencing and *de novo* assembly of a chloroplast genome. *BMC Genomics* **14**, doi: 10.1186/1471-2164-14-670 (2013).
19. Fichot, E. B. & Norman, R. S. Microbial phylogenetic profiling with the Pacific Biosciences sequencing platform. *Microbiome* **1**, doi: 10.1186/2049-2618-1-10 (2013).
20. Chen, J. *et al.* High-throughput platform for real-time monitoring of biological processes by multicolor single-molecule fluorescence. *Proc Natl Acad Sci USA* **111**, 664–669 (2014).
21. Nakano, K. *et al.* First complete genome sequence of *Pseudomonas aeruginosa* (Schroeter 1872) Migula 1900 (DSM 50071<sup>T</sup>), determined using PacBio single-molecule real-time technology. *Genome Announc* **3**, e00932–00915 (2015).
22. Bendich, A. J. Structural analysis of mitochondrial DNA molecules from fungi and plants using moving pictures and pulsed-field gel electrophoresis. *J Mol Biol* **255**, 564–588 (1996).
23. Cuenca, A., Petersen, G. & Seberg, O. The complete sequence of the mitochondrial genome of *Butomus umbellatus*—a member of an early branching lineage of Monocotyledons. *PLoS One* **8**, e61552 (2013).
24. Tamura, K. *et al.* MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* **28**, 2731–2739 (2011).
25. Raubeson, L. A. *et al.* Comparative chloroplast genomics: analyses including new sequences from the angiosperms *Nuphar advena* and *Ranunculus macranthus*. *BMC Genomics* **8**, doi: 10.1186/1471-2164-8-174 (2007).
26. Sloan, D. B. & Taylor, D. R. Testing for selection on synonymous sites in plant mitochondrial DNA: the role of codon bias and RNA editing. *J Mol Biol* **70**, 479–491 (2010).
27. Morton, B. R. & Levin, J. A. The atypical codon usage of the plant psbA gene may be the remnant of an ancestral bias. *Proc Natl Acad Sci USA* **94**, 11434–11438 (1997).
28. Gray, M. W. Evolution of organellar genomes. *Curr Opin Genet Dev* **9**, 678–687 (1999).
29. Nakazono, M. *et al.* The rps3-rpl16-nad3-rps12 gene cluster in rice mitochondrial DNA is transcribed from alternative promoters. *Curr Genet* **27**, 184–189 (1995).
30. Sugiyama, Y. *et al.* The complete nucleotide sequence and multipartite organization of the tobacco mitochondrial genome: comparative analysis of mitochondrial genomes in higher plants. *Mol Genet Genomics* **272**, 603–615 (2005).
31. Alverson, A. J. *et al.* Insights into the evolution of mitochondrial genome size from complete sequences of *Citrullus lanatus* and *Cucurbita pepo* (Cucurbitaceae). *Mol Biol Evol* **27**, 1436–1448 (2010).
32. Turmel, M., Otis, C. & Lemieux, C. The complete mitochondrial DNA sequence of *Mesostigma viride* identifies this green alga as the earliest green plant divergence and predicts a highly compact mitochondrial genome in the ancestor of all green plants. *Mol Biol Evol* **19**, 24–38 (2002).
33. Bremer, B. *et al.* An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. *Bot J Linn Soc* **161**, 105–121 (2009).
34. Lynn, D. J. *et al.* Bioinformatic discovery and initial characterisation of nine novel antimicrobial peptide genes in the chicken. *Immunogenetics* **56**, 170–177 (2004).
35. Mower, J. P., Sloan, D. B. & Alverson, A. J. Plant mitochondrial genome diversity: the genomics revolution. In *Plant Genome Diversity*. Vol. 1 (eds J. F., Wendel *et al.*) 123–144 (Springer-Verlag Wien, 2012).
36. Adams, K. L., Qiu, Y. L., Stoutemyer, M. & Palmer, J. D. Punctuated evolution of mitochondrial gene content: high and variable rates of mitochondrial gene loss and transfer to the nucleus during angiosperm evolution. *Proc Natl Acad Sci USA* **99**, 9905–9912 (2002).
37. Richardson, A. O., Rice, D. W., Young, G. J., Alverson, A. J. & Palmer, J. D. The “fossilized” mitochondrial genome of *Liriodendron tulipifera*: ancestral gene content and order, ancestral editing sites, and extraordinarily low mutation rate. *BMC Biol* **11**, doi: 10.1186/1741-7007-11-29 (2013).
38. Shields, D. C. & Wolfe, K. H. Accelerated evolution of sites undergoing mRNA editing in plant mitochondria and chloroplasts. *Mol Biol Evol* **14**, 344–349 (1997).
39. Mower, J. P. Modeling sites of RNA editing as a fifth nucleotide state reveals progressive loss of edited sites from angiosperm mitochondria. *Mol Biol Evol* **25**, 52–61 (2008).
40. Stoltzfus, A. On the possibility of constructive neutral evolution. *J Mol Biol* **49**, 169–181 (1999).
41. Kreunen, S. S. & Osborn, J. M. Pollen and anther development in *Nelumbo* (Nelumbonaceae). *Am J Bot* **86**, 1662–1676 (1999).
42. Liu, Y. *et al.* The mitochondrial genome of the lycophyte *Huperzia squarrosa*: the most archaic form in vascular plants. *PLoS One* **7**, e35168 (2012).
43. Turmel, M., Otis, C. & Lemieux, C. The mitochondrial genome of *Chara vulgaris*: insights into the mitochondrial DNA architecture of the last common ancestor of green algae and land plants. *Plant Cell* **15**, 1888–1903 (2003).
44. Zhu, A., Guo, W., Jain, K. & Mower, J. P. Unprecedented heterogeneity in the synonymous substitution rate within a plant genome. *Mol Biol Evol* **31**, 1228–1236 (2014).
45. Kitazaki, K. *et al.* A horizontally transferred tRNACys gene in the sugar beet mitochondrial genome: evidence that the gene is present in diverse angiosperms and its transcript is aminoacylated. *Plant J* **68**, 262–272 (2011).
46. Knie, N., Polsakiewicz, M. & Knoop, V. Horizontal gene transfer of chlamydial-like tRNA genes into early vascular plant mitochondria. *Mol Biol Evol* **32**, 629–634 (2015).
47. Chaw, S. M., Shih, A. C. C., Wang, D., Wu, Y. W. & Liu, S. M. The mitochondrial genome of the gymnosperm *Cycas taitungensis* contains a novel family of short interspersed elements, Bpu sequences, and abundant RNA editing sites. *Mol Biol Evol* **25**, 603–615 (2008).
48. Kent, W. J. BLAT—the BLAST-like alignment tool. *Genome Res* **12**, 656–664 (2002).
49. Miller, J. R. *et al.* Aggressive assembly of pyrosequencing reads with mates. *Bioinformatics* **24**, 2818–2824 (2008).
50. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

51. Harris, R. S. *Improved pairwise alignment of genomic DNA*. PhD thesis, Pennsylvania State Univ. (2007).
52. Schattner, P., Brooks, A. N. & Lowe, T. M. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Res* **33**, W686–W689 (2005).
53. Conant, G. C. & Wolfe, K. H. GenomeVx: simple web-based creation of editable circular chromosome maps. *Bioinformatics* **24**, 861–862 (2008).
54. Thiel, T., Michalek, W., Varshney, R. & Graner, A. Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). *Theor Appl Genet* **106**, 411–422 (2003).
55. McGinnis, S. & Madden, T. L. BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res* **32**, W20–W25 (2004).
56. Tang, H. *et al.* Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. *Genome Res* **18**, 1944–1954 (2008).
57. Gambino, G., Perrone, I. & Gribaudo, I. A rapid and effective method for RNA extraction from different tissues of grapevine and other woody plants. *Phytochem Anal* **19**, 520–525 (2008).
58. Larkin, M. A. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947–2948 (2007).
59. Notsu, Y. *et al.* The complete sequence of the rice (*Oryza sativa* L.) mitochondrial genome: frequent DNA sequence acquisition and loss during the evolution of flowering plants. *Mol Genet Genomics* **268**, 434–445 (2002).
60. Posada, D. jModelTest: phylogenetic model averaging. *Mol Biol Evol* **25**, 1253–1256 (2008).
61. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
62. Guindon, S. & Gascuel, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* **52**, 696–704 (2003).
63. Paradis, E., Claude, J. & Strimmer, K. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**, 289–290 (2004).
64. Yu, G., Smith, D., Zhu, H., Guan, Y. & Lam, T. T. Y. *ggtree: a phylogenetic tree viewer for different types of tree annotations*, Available at: <https://github.com/Bioconductor-mirror/ggtree/tree/release-3.2> (Accessed: 12 April 2016) (2015).
65. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**, 1586–1591 (2007).
66. Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome Res* **19**, 1639–1645 (2009).

## Acknowledgements

The authors thank Professor Palmer JD (Department of Biology, Indiana University) for direction on mtDNA isolation, and Jiajian Zhou (BSgenomics Co., Ltd., ShenZhen) for bioinformatic help. This research is financially supported by National Natural Science Foundation of China (31271310).

## Author Contributions

Designed the general idea and experiments of the study: Y.D., Z.W. Isolated the *N. nucifera* mtDNA and carried out most of the experiments, and drafted the manuscript: S.G., Z.W. Performed annotation of mtDNA and some of the experiments: H.Z., Y.Z., Z.Z. and D.L. Modified the final manuscript: S.G., Z.W. and Y.D.

## Additional Information

**Supplementary information** accompanies this paper at <http://www.nature.com/srep>

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

**How to cite this article:** Gui, S. *et al.* The mitochondrial genome map of *Nelumbo nucifera* reveals ancient evolutionary features. *Sci. Rep.* **6**, 30158; doi: 10.1038/srep30158 (2016).



This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>

© The Author(s) 2016