The first complete mitochondrial genome of eggplant (Solanum melongena)

Dandan Li, Weiliu Li, Guiyun Gan, Wenjia Li, Yaqin Jiang, Xuyu Liang, Riyuan Chen and Yikui Wang

Institute of Vegetable Research, Guangxi Academy of Agricultural Sciences, Nanning, China;

ABSTRACT

Eggplant is an important vegetable crop because of its rich nutrition, but to date no mitochondrial genome has been reported. In this study, the complete mitochondrial genome of the eggplant was sequenced. The complete mitochondrial genome was 498,136bp, linear structure, containing 54 protein-coding genes, four rRNAs, and 32 tRNAs. The phylogenetic tree supported the hypothesis that the eggplant is most closely related to Solanum tuberosum and Solanum lycopersicum.

Solanum is one of the largest and the most economically important family in the Solanaceae, which encompasses more than 1500 species (Bohs and Olmstead, 1997) and is widely distributed in the tropical and subtropical regions of the world (Whalen 1984; Levin et al. 2006; Dupin et al. 2017). Eggplant is an important Solanaceae crop; however, when compared with other Solanaceae plants (such as tobacco, potato, tomato and pepper), the research work of eggplant is relatively backward. Now, we reported the complete mitochondrial genome of Solanum melongena, which is based on the next-generation sequence. We believe that our study provides the fundamental information for unraveling the evolution and domestication of the eggplant and may ultimately lead to further improvement of Solanaceae crops.

Here, we present the complete mitochondrial genome of eggplant collected from the Vegetable Research Institute, Guangxi Academy of Agricultural Science (28°N and 118°E), Guangxi Province, China. The identification was confirmed by YiKui Wang. The material was deposited at the Seed Bank of Guangxi Academy of Agricultural Science (accession number: 177).

DNA extraction, genome sequencing, assembly, and genome analysis

Approximately 5 g of fresh leaves was harvested for mtDNA isolation using an improved extraction method (Chen et al. 2011). After DNA isolation, 1 μg of purified DNA was fragmented to construct short-insert libraries (insert size 430 bp) according to Illumina’s instructions, then sequenced on the Illumina Hiseq 4000 (Erik et al. 2011). The high molecular weight DNA was purified and used for PacBio library preparation, Blue Pippin size selection, then sequenced on the Sequel Sequencer. Prior to assembly, Illumina raw reads were filtered firstly. This filtering step was performed in order to remove the reads with adaptors, the reads showing a quality score below 20 (Q < 20), the reads containing a percentage of uncalled based (‘N’ characters) equal or greater than 10% and the duplicated sequences. The mitochondrial genome was reconstructed using a combination of Pacbio Sequel and the Illumina Hiseq data, and the following three steps were used to assemble mitochondria genomes. First, Assemble the genome framework by the both Illumina and Pacbio data using SPAdesv3.10.1 (Antipov et al., 2016). Secondly, verifying the assembly and completing the circle or linear characteristic of the mitochondria genome, filling gaps if there were. Third, clean reads were mapped to the assembled mitochondria genome to correct the wrong bases, judge if there is any insertion and deletion.

The mitochondrial genes were annotated using homology alignments and de novo prediction, and the Evidence Modeler/v1.1.1 was used to integrate the gene set (Haas et al. 2008). Transfer RNA (tRNA) genes and ribosome RNA (rRNA) genes were predicted by tRNA scan-SE (Lowe and Eddy 1997) and rRNA mmer 1.2 (Lagesen et al. 2007). Then, Blast (Altschul et al. 1990) search (E-value ≤ 1e−5, minimal alignment length percentage ≥ 40%) gene sequence against 5 databases for gene function annotation was used. The five databases are: KEGG (Knehisa; 1997; Kanehisa et al. 2004, 2006) (Kyoto Encyclopedia of Genes and Genomes), COG (Tatusov et al. 1997, 2003) (Clusters of Orthologous Groups), NR (Non-Redundant Protein Database databases), Swiss-Prot (Magrane 2011), and GO (Ashburner et al. 2000) (Gene Ontology).

The complete mitochondrial genomes of other Solanaceae plants were downloaded from NCBI. ClustalW was used to align the mtDNA sequences under default parameters (Larkin
et al. 2007), and the alignment was checked manually. The maximum-likelihood (ML) methods were performed for the genome-wide phylogenetic analyses using PhyML3.0 (Guindon et al. 2010). Nucleotide substitution model selection was estimated with j Model Test 2.1.10 (Darriba et al. 2012) and Smart Model Selection in PhyML 3.0. The model GTR + G was selected for ML analyses with 1000 bootstrap replicates to calculate the bootstrap values (BS) of the topology. The results were treated with iTOL 3.4.3 (Letunic and Bork 2016).

**Conclusions**

The complete mitochondrial genome was 498,136 bp (linear structure, two contigs) and the GC content was 43.9%. There were 54 protein genes, 32 tRNAs and 4 rRNAs annotated. The percentage of three type of gene length is 8.28%, 1.01% and 0.49%, respectively. Through data analysis, we also found 264 SSRs and 285 edit sites. From the constructed phylogenetic tree, we use complete mitogenome sequences (Figure 1). The phylogeny tree supported the assertion that the eggplant is most closely related to *Solanum tuberosum* and *Solanum lycopersicum*.

**Disclosure statement**

No potential conflict of interest was reported by the author(s)

**Funding**

This work was supported by the National Natural Science Foundation of China (31660573), Guangxi Natural Science Foundation (2018GXNSFDA050015 ) and Science and Technology Major Project of Guangxi (GuiKeAA20108005-3).

**Data availability statement**

The genome sequence data that support the findings of this study are openly available in GenBank of NCBI at [https://www.ncbi.nlm.nih.gov/](https://www.ncbi.nlm.nih.gov/) under the accession MT122986 and MT122987. The associated SRA number is SAMN16746491.

**References**

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol. 215(3):403–410.

Antipov D, Korobeynikov A, Mclean JS, Pevzner PA. 2016. hybridSPAdes: an algorithm for hybrid assembly of short and long reads. Bioinformatics. 32(7):1009–1015.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. 2000. Gene ontology: tool for the unification of biology. Nat Genet. 25(1):25–29.

Bohs L, Olmstead RG. 1997. Phylogenetic relationships in *Solanum* (Solanaceae) based on ndhF sequences. Systemat Bot. 22(1):5–17.

Chen J, Rongzhan G, Shengxin C, Tongqing D, Hongsheng Z, Han X, Xiaoyu Z. 2011. Substoichiometrically different mitotypes coexist in mitochondrial genomes of *Brassica napus* L. PloS One. 6(3):e17662.

Darriba D, Taboada GL, Doallo R, Posada D. 2012. jModelTest 2: more models, new heuristics and parallel computing. Nat Methods. 9(8):772.

Dupin J, Matzke NJ, Säkinen T, Knapp S, Olmstead RG, Bohs L, Smith SD. 2017. Bayesian estimation of the global biogeographical history of the Solanaceae. J Biogeogr. 44(4):887–899.

Erik BM, Sverker L, Joakim L. 2011. Large scale library generation for high throughput sequencing. PLOS One. 6:231–238.
Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 59(3):307–321.

Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR. 2008. Automated eukaryotic gene structure annotation using EVIDencEModeler and the Program to assemble spliced alignments. Genome Biol. 9(1):R7.

Knehisa M. 1997. A database for post-genome analysis. Trends Genet. 9:375–376.

Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. 2004. The KEGG resource for deciphering the genome. Nucleic Acids Res. 32(Database issue):D277–D280.

Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M. 2006. From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res. 34(Database issue):D354–D357.

Lagesen K, Hallin P, Rødland EA, Staerfeldt H-H, Rognes T, Ussery DW. 2007. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res. 35(9):3100–3108.

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, et al. 2007. Clustal W and clustal X version 2.0. Bioinformatics. 23(21):2947–2948.

Letunic I, Bork P. 2016. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res. 44(W1):W242–W245.

Levin RA, Myers NR, Bohs L. 2006. Phylogenetic relationships among the,"spiny solanums" (Solanum subgenus Leptostemonum, Solanaceae). Am J Bot. 93(1):157–169.

Lowe TM, Eddy SR. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 25(5):955–964.

Magrane M. 2011. UniProt Knowledgebase: a hub of integrated protein data. Database (Oxford). 2011:bar009.

Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, et al. 2003. The COG database: an updated version includes eukaryotes. BMC Bioinformatics. 4:41–41.

Tatusov RL, Koonin EV, Lipman DJ. 1997. A genomic perspective on protein families. Science. 278(5338):631–637.

Whalen MD. 1984. Conspectus of species groups in Solanum subgenus Leptostemonum. Gentes Herbarum. 12(4):179.