

The complete plastid genomes of four species from Brassicales

Weixue Mu^{1,2}, Ting Yang^{1,2}, Xin Liu^{1,2#}

1. BGI-Shenzhen, Shenzhen 518083, China

2. China National GeneBank, BGI-Shenzhen, Shenzhen 518120, China

#Correspondence: Xin Liu (liuxin@genomics.cn)

ABSTRACT

Brassicales is a diverse angiosperm order with about 4,700 recognized species. Here, we assembled and described the complete plastid genomes from four species of Brassicales: *Capparis urophylla* F.Chun (Capparaceae), *Carica papaya* L. (Caricaceae), *Cleome rutidosperma* DC. (Cleomaceae), and *Moringa oleifera* Lam. (Moringaceae), including two plastid genomes newly assembled for two families (Capparaceae and Moringaceae). The four plastid genomes are 159,680 base pairs on average in length and encode 78 protein-coding genes. The genomes each contains a typical structure of a Large Single-Copy (LSC) region and a Small Single-Copy (SSC) region separated by two Inverted Repeat (IR) regions. We performed the maximum-likelihood (ML) phylogenetic analysis using three different data sets of 66 protein-coding genes (ntAll, ntNo3rd and AA). Our phylogenetic results from different dataset are congruent, and are consistent with previous phylogenetic studies of Brassicales.

KEYWORDS Brassicales; plastid genome; phylogenetic analysis

Brassicales, an order of flowering plants that contains ~4,700 species (Kiefer et al. 2014; Cardinal - McTeague et al. 2016), is a monophyletic group consisting 17 families including several model families, such as Brassicaceae, for development and evolutionary biological studies (Soltis et al. 2011; Chase et al. 2016). Brassicales are

known for their rich diversities in morphology, physiology, development, and chemical traitse (Koenig and Weigel, 2015). Here, we reported the complete plastid genomes of four Brassicales species that were assembled using the Next-Generation Sequencing approach with BGISEQ-500 (BGI, Shenzhen) paired-end reads.

In this study, the samples were collected from Ruili Botanical Garden at Yunnan, China. The voucher specimens have been deposited in the Herbarium of China National GeneBank (HCNGB). The fresh leaves of each sample were used to extract total genomic DNA using the modified CTAB method (Doyle and Doyle 1987). High-throughput sequencing was carried out using an BGISEQ-500 platform. Approximately 50 Gb high quality, 100 base pairs (bp) paired-end reads were obtained for each sample. The raw reads with low quality were filtered out using the program SOAPfilter_v2.2. *De novo* assemblies were carried out using the seed-extension-based *de novo* assembler NOVOPlasty v2.5.9 (Dierckxsens et al. 2016). A complete *rbcL* gene sequence of *Arabidopsis thaliana* (downloaded from NCBI, accession number U91966) was used as the seed to conduct the assemblies. The longest contig of each sample were further blasted against the chloroplast database (downloaded from NCBI, including 2,503 non-redundant species), and the resulted best-hit sequences (minimum requirement: e-value < 10⁻⁷ and identity>95%) were used as the reference for further assembly using an ‘baiting and iterative mapping’ approach assembler MITObim v1.8 (Hahn et al. 2013). Complete plastid genomes were recovered for all the four species. Plastid genomes were then annotated on the standard web-based program GeSeq (Tillich et al. 2017), and the annotation of protein-coding genes were further conducted with GeneWise v2.4.1 (Birney et al. 2000). The complete plastid genomes have been submitted to the CNGB Nucleotide Sequence Archive (CNSA : <https://db.cngb.org/cnsa>; accession number: CNA0000813 (*Capparis urophylla* F.Chun); CNA0000814 (*Carica papaya* L.); CNA0000815 (*Cleome rutidosperma* DC.); CNA0000816 (*Moringa oleifera* Lam.).)

The assembled four Brassicales plastid genomes ranged from 155,359 bp (*Capparis urophylla* F.Chun) to 163,131 bp (*Moringa oleifera* Lam.) in length, and all share a typical quadripartite structure: two inverted repeats (IRs), a large single-copy region (LSC), and a small single-copy region (SSC). As for *Carica papaya* L. and *Cleome rutidosperma* DC., the total length of the plastid genomes are 161,752 bp and 158,477 bp, respectively. The plastid genome of *Carica papaya* L. has the longest LSC region of 130,436 bp, along with two 3,483 bp IR regions and a shortest 24,350 bp SSC region, while the plastid genome of *Capparis urophylla* F. Chun has the shortest LSC (84,518 bp), two IRs (10,318 bp) and a longest SSC (50,205 bp), respectively. The length of LSC, IR and SSC regions of *Cleome rutidosperma* DC. and *Moringa oleifera* Lam. are 107,763 bp, 2,888 bp, 45,028 bp and 102,342 bp, 3,710 bp, 48,715 bp, respectively. All four species have the same number of 78 protein-coding genes, four ribosomal RNAs, while the number of transfer RNAs varies from 36 (*Cleome rutidosperma* DC.), 37 (*Carica papaya* L. and *Moringa oleifera* Lam.) to 38 (*Capparis urophylla* F.Chun).

We performed phylogenetic analyses with all four newly assembled plastid genome sequences together with 16 published plastid genomes of Brassicales species and two Malvales species. The plastid sequences were aligned with MAFFT V7.017 (Katoh et al. 2013) on the base following three data sets: (1) all nucleotide positions (ntAll); (2) the first and second codon positions (ntNo3rd); (3) the amino acid (AA) of 78 protein-coding genes (Ruhfel et al. 2014). The Maximum-Likelihood trees were constructed under the GTRCAT (PROTGAMMAWAG for AA) model using the parallel version of RAxML v8.2.4 (Stamatakis 2014) with 100 bootstraps. The phylogenetic results revealed that the relationships constructed using the three different data sets were highly consistent at the family level (Figure 1). The plastid phylogeny recovered in our study is congruent with the previous phylogenetic study of Brassicales (APG 2016).

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the grants of Basic Research Program, the Shenzhen Municipal Government, China (No.JCYJ20150529150505656) and (No.JCYJ20150831201643396), as well as the funding to State Key Laboratory of Agricultural Genomics (No.2011DQ782025), and Guangdong Provincial Key Laboratory of Genome Read and Write (No.2017B030301011)

References

- APG, 2016. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. *Bot. J. Linn. Soc.* 181, 1–20. 0.
- Birney E, Durbin R. 2000. Using GeneWise in the Drosophila annotation experiment. *Genome Research.* 10:547-548.
- Cardinal - McTeague, W. M., K. J. Sytsma, and J. C. Hall. 2016. Biogeography and diversification of Brassicales: A 103 million year tale. *Molecular Phylogenetics and Evolution* 99: 204–224.
- Chase, M. W., M. J. M. Christenhusz, M. F. Fay, J. W. Byng, W. S. Judd, D. E. Soltis, D. J. Mabberley, et al. 2016. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. *Botanical Journal of the Linnean Society* 181: 1–20.
- Dierckxsens N, Mardulyn P, Smits G. 2016. NOVOPlasty: de novo assembly of organelle genomes from whole genome data. *Nucleic acids research.* 45:e18-e18.
- Doyle JJ, Doyle JL. 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem Bull.* 19:11–15.

Hahn C, Bachmann L, Chevreux B. 2013. Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads-a baiting and iterative mapping approach. *Nucleic Acids Research*. 41:e129-e129.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*. 30:772-780.

Kiefer, M., R. Schmickl, D. A. German, T. Mandáková, M. A. Lysak, I. A. Al - Shehbaz, A. Franzke, et al. 2014. BrassiBase: introduction to a novel knowledge database on Brassicaceae evolution. *Plant & Cell Physiology* 55: e3.

Koenig, D., and D. Weigel. 2015. Beyond the thale: comparative genomics and genetics of *Arabidopsis* relatives. *Nature Reviews. Genetics* 16: 285–298.

Soltis, D. E., S. A. Smith, N. Cellinese, K. J. Wurdack, D. C. Tank, S. F. Brockington, N. F. Refulio - Rodriguez, et al. 2011. Angiosperm phylogeny: 17 genes, 640 taxa. *American Journal of Botany* 98: 704–730.

Ruhfel BR, Gitzendanner MA, Soltis PS, Soltis DE, Burleigh JG. 2014. From algae to angiosperms-inferring the phylogeny of green plants (Viridiplantae) from 360 plastid genomes. *BMC Evolutionary Biology*. 14:23.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*. 30:1312-1313.

Tillich M, Lehwark P, Pellizzer T, Ulbricht-Jones ES, Fischer A, Bock R, Greiner S. 2017. GeSeq-versatile and accurate annotation of organelle genomes. *Nucleic Acids Research*. 45:W6-W11.

Figure legend

Figure 1. Maximum-likelihood phylogenetic tree based on 22 ntAll data sets. The numbers on the nodes indicate bootstrap support values.

