Complete Chloroplast Genome Sequence and Annotation of the Tropical *japonica* Group of Asian Cultivated Rice (*Oryza sativa* L.)

Shuo Wang,[a,b] Li-zhi Gao[a,b]

Faculty of Environmental Science and Engineering[a] and Faculty of Life Science and Technology,[b] Kunming University of Science and Technology, Kunming, China

We announce here the first complete chloroplast genome sequence of the tropical *japonica* rice, along with its genome structure and functional annotation. The plant was collected from Indonesia and deposited as a germplasm accession of the International Rice GenBank Collection (IRGC 66630) at the International Rice Research Institute (IRRI). This genome provides valuable data for the future utilization of the germplasm of rice.

Rice (*Oryza sativa* L.), as an important staple crop of the family *Poaceae*, is distributed widely across diverse tropical-to-temperate regions of both hemispheres and provides the vast majority of daily caloric intake for half the world's population. It is also well known for its great genetic diversity within species (1, 2), which can be categorized into five distinct varietal groups: *indica*, *aus/boro*, *aromatic* (basmati/sadri), temperate *japonica*, and tropical *japonica* (alias *javanica*) (3–6). Along with the arrival of genome sequencing era, the *indica* and temperate *japonica* rice became some of the first few crop species having their nuclear, chloroplast, and mitochondrial genomes completely sequenced (7–12), which consolidated their valuable status as model systems for other grass species.

The tropical *japonica* rice (*O. sativa* subsp. tropical *japonica*) was most likely domesticated in northern parts of Southeast Asia or South China, and then introduced southward to Southeast Asia and from there to West Africa and Latin American countries (13). Here we first announce its complete chloroplast genome sequence and functional annotations. The plant was collected from Indonesia and deposited as a germplasm accession of the International Rice Collection (IRGC 66630) at the International Rice Research Institute (IRRI), which was then prepared for sequencing as an accession of the 3,000 Rice Genomes Project (3K RGP [14]). Genomic DNA was extracted by a modified cetyltrimethylammonium bromide (CTAB) method at IRRI and shipped to BGI-Shenzhen for the construction of Illumina index libraries. After quality control, at least 3 μg of genomic DNA was randomly fragmented by sonication and size fractionated by electrophoresis. DNA fragments of approximately 500 bp in length were then purified, labeled with 6-bp nucleotide multiplex identifiers, and followed by pooling prior to library construction for next-generation sequencing (NGS). Each library was sequenced in six or more lanes on the Illumina HiSeq 2000 platform to generate 90-bp paired-end reads. The reads were subsequently extracted based on the corresponding unique nucleotide multiplex identifiers and filtered by deleting those with adapter contamination or containing >50% low-quality bases (quality value, ≤5) (14).

Approximately 1.5 Gb of Illumina paired-end total DNA sequencing data (GigaScience database [15]) were filtered with NCBI-blast version 2.2.31+ (ftp://ftp.ncbi.nih.gov/blast/) to obtain chloroplast DNA reads. The filtered chloroplast DNA reads were then subjected to SOAP*denovo*2 (16), ABySS version 1.9.0 (17), and SPAdes version 3.1.0 (18) for several runs of assembly. The final assembly resulted in a complete circular genome sequence with a length of 134,536 bp and G+C content of 39.7%. Annotation was performed with Dual Organellar GenoMe Annotator (DOGMA [19]) using default parameters to predict protein-coding genes, tRNA genes, and ribosomal RNA (rRNA) genes. For genes with low sequence identity, manual annotation was performed to determine the positions of start and stop codons depending on the translated amino acid sequence using the chloroplast/bacterial genetic code.

Nucleotide sequence accession numbers. The source accession for this DNA sample at IRRI is IRGC 66630 (http://iris.irri.org/germplasm2/id/365929), and the accession number for the genetic stock attached to the DNA sample is IRGC 126310 (http://iris.irri.org/germplasm2/id/3059145). The complete chloroplast genome sequence with all genes annotated has been submitted to GenBank under the accession number KT289404.

REFERENCES

1. **Li ZK, Rutger JN**. 2000. Geographic distribution and multilocus organization of isozyme variation of rice (*Oryza sativa* L.). Theor Appl Genet **101**:379–387. http://dx.doi.org/10.1007/s001220051494.
2. **Yu SB, Xu WJ, Vijayakumar CHM, Ali J, Fu BY, Xu JL, Marghirang R, Domingo J, Jiang YZ, Aquino C, Virmani SS, Li ZK**. 2003. Molecular diversity and multilocus organization of the parental lines used in the International Rice Molecular Breeding Program. Theor Appl Genet **108**:131–140. http://dx.doi.org/10.1007/s00122-003-1400-3.
3. **Vaughan DA**. 1994. The wild relatives of rice: a genetic resources handbook, p 66. International Rice Research Institute, Manila, Philippines.
4. **Izawa T**. 2008. The process of rice domestication: a new model based on recent data. Rice **1**:127–134. http://dx.doi.org/10.1007/s12284-008-9014-7.
5. **Kovach MJ, Sweeney MT, McCouch SR**. 2007. New insights into the history of rice domestication. Trends Genet **23**:578–587. http://dx.doi.org/10.1016/j.tig.2007.08.012.
6. **Garris AJ, Tai TH, Coburn J, Kresovich S, McCouch S**. 2005. Genetic

structure and diversity in *Oryza sativa* L. Genetics **169:**1631–1638. http://dx.doi.org/10.1534/genetics.104.035642.

7. **Yu J, Hu S, Wang J, Wong GK, Li S, Liu B, Deng Y, Dai L, Zhou Y, Zhang X, Cao M, Liu J, Sun J, Tang J, Chen Y, Huang X, Lin W, Ye C, Tong W, Cong L, Geng J, Han Y, Li L, Li W, Hu G, Huang X, Li W, Li J, Liu Z, Li L, Liu J, Qi Q, Liu J, Li L, Li T, Wang X, Lu H, Wu T, Zhu M, Ni P, Han H, Dong W, Ren X, Feng X, Cui P, Li X, Wang H, Xu X, Zhai W, Xu Z, et al.** 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). Science **296:**79–92. http://dx.doi.org/10.1126/science.1068037.

8. **Goff SA, Ricke D, Lan TH, Presting G, Wang R, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H, Hadley D, Hutchison D, Martin C, Katagiri F, Lange BM, Moughamer T, Xia Y, Budworth P, Zhong J, Miguel T, Paszkowski U, Zhang S, Colbert M, Sun WL, Chen L, Cooper B, Park S, Wood TC, Mao L, Quail P, Wing R, Dean R, Yu Y, Zharkikh A, Shen R, Sahasrabudhe S, Thomas A, Cannings R, Gutin A, Pruss D, Reid J, Tavtigian S, Mitchell J, Eldredge G, Scholl T, Miller RM, Bhatnagar S, Adey N, Rubano T, Tusneem N, et al.** 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. japonica). Science **296:**92–100. http://dx.doi.org/10.1126/science.1068275.

9. **Tang J, Xia H, Cao M, Zhang X, Zeng W, Hu S, Tong W, Wang J, Wang J, Yu J, Yang H, Zhu L.** 2004. A comparison of rice chloroplast genomes. Plant Physiol **135:**412–420. http://dx.doi.org/10.1104/pp.103.031245.

10. **Tian X, Zheng J, Hu S, Yu J.** 2006. The rice mitochondrial genomes and their variations. Plant Physiol **140:**401–410. http://dx.doi.org/10.1104/pp.105.070060.

11. **Shimda H, Sugiuro M.** 1991. Fine structural features of the chloroplast genome: comparison of the sequenced chloroplast genomes. Nucleic Acids Res **19:**983–995. http://dx.doi.org/10.1093/nar/19.5.983.

12. **Notsu Y, Masood S, Nishikawa T, Kubo N, Akiduki G, Nakazono M, Hirai A, Kadowaki K.** 2002. 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. http://dx.doi.org/10.1007/s00438-002-0767-1.

13. **Khush GS.** 1997. Origin, dispersal, cultivation and variation of rice p 25–34. *In* Sasaki T, Moore G (ed) Oryza: From molecule to plant. Springer International Publishing, The Netherlands.

14. **The 3,000 Rice Genomes Project.** 2014. The Rice 3,000 Genome Project. GigaScience **3:**e7. http://dx.doi.org/10.1186/2047-217X-3-7.

15. **The 3,000 Rice Genomes Project.** 2014. The Rice 3,000 Genome Project. GigaScience Database. http://dx.doi.org/10.5524/200001.

16. **Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, Tang J, Wu G, Zhang H, Shi Y, Liu Y, Yu C, Wang B, Lu Y, Han C, Cheung DW, Yiu SM, Peng S, Xiaoqian Z, Liu G, Liao X, Li Y, Yang H, Wang J, Lam TW, Wang J.** 2012. SOAP*denovo*2: an empirically improved memory-efficient short-read *de novo* assembler. GigaScience **1:**e18. http://dx.doi.org/10.1186/2047-217X-1-18.

17. **Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, Birol I.** 2009. ABySS: a parallel assembler for short read sequence data. Genome Res **19:**1117–1123. http://dx.doi.org/10.1101/gr.089532.108.

18. **Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA.** 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol **19:**455–477. http://dx.doi.org/10.1089/cmb.2012.0021.

19. **Wyman SK, Jansen RK, Boore JL.** 2004. Automatic annotation of organellar genomes with DOGMA. Bioinformatics **20:**3252–3255. http://dx.doi.org/10.1093/bioinformatics/bth352.