Bangladesh J. Bot. 50(3): 895-901, 2021 (September) Special DOI: https://doi.org/10.3329/bjb.v50i0.56442

COMPLETE CHLOROPLAST GENOME OF *ORYZA SATIVA* L. (B810S) AND ITS PHYLOGENETIC ANALYSIS

KEBAO SONG, CONGTIAN WANG*, ZHONGBO LI AND PENG NING

College of Environment and Biology, HuaiHua Vocational and Technical College, Huaihua, Hunan Province418000, PR China

Keywords: *Oriza Sativa* L. (B810S), Chloroplast genome, Phylogenetic analysis, China

Abstract

The complete chloroplast (cp) genome of *Oryza sativa* L.**(**B810S**)** was 134546 bp in length in the study, which contains 149 genes including 99 coding protein genes, 41 transfer RNA genes, 8 ribosomal RNA genes and 1 non-coding region by gene annotation. A total of 20879 amino acids were encoded by this cp genome, TTT (Phe) and TTG (Leu) codon were the most frequent amino acids, whereas the ACC (Thr), GCC (Ala), CTC (Leu), and AAC (Asn) codon were the least frequent ones. The content of the four bases on the cp genome were 30.6% for A, 30.4% for T, 19.4% for C and 19.6% for G, respectively. Obviously, the A+T (61.0%) content is more higher than G+C (39.0%). The gene order and content are the same as those of previously reported cp genome of Rice. Phylogenetic analysis was implemented based on concatenated amino acid sequences of 99 protein-coding genes using Neighbor-Joining method (NJ) method. Therefore, the complete B810S cp genome provides interesting insights and valuable information that can be used to identify related species and reconstruct its phylogeny.

Introduction

Rice, 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 (Group II, 2012). Rice is divided into five distinct varietal groups: *indica*, *aus/boro*, *aromatic* (basmati/sadri), temperate *japonica*, and tropical *japonica* (alias *javanica*) (Izawa 2008, Kovach *et al*. 2007). At present, more than 20 species of rice from genus *Oryza* were found in the world, but only two species of the genus *Oryza*---*O. sativa* and O. *glaberrima* are cultivated (Evenson and Gollin 1997, Sang and Ge 2007). According to many related studies, there are great genetic diversity within species of rice (Li and Rutger 2000). About one half of the species in *Oryza* genus are allotetraploids that originated through inter-specific hybridization and genome doubling (Bao and Ge 2008, Jacquemin *et al.* 2013). The *Oryza* genus with an AA genome type, is one of the most important species, which are most likely domesticated in northern parts of Southeast Asia and South China as a major food crop (Wang et al. 2016). People have spent much attention to understand the genetic makeup and phylogeny of the *Oryza* genus (Asaf *et al.* 2017). Who used the complete cp genome to analysis the phylogenetic relationships of wild rice and other rices (Asaf *et al.* 2017). Presently, although many cp genomes of Rice from *Oryza* genus were in Gene Bank, there are a little available information about the cp genomes of B810S from China.

The chloroplast genome is a typically circular DNA molecule, the rise ranging from 39.4 to 200.8 kb among photosynthetic plant species (Turmel *et al.* 1999), and is maternally inherited and possesses its own genome encoding many chloroplast specific components (Palmer *et al.* 1988, Sugiura 1989). With the development of high throughput sequencing technologies and the conserved features of chloroplast genomes, over 1,000 species in Viridiplantae have been completely sequenced (Tong *et al.* 2016). For example, Hiratsuka *et al.* (1989) reported that the

*Author for correspondence: <18390913065@163.com>.

chloroplast genome of rice Nipponbare (O. sativa L. ssp. japonica) have been 134,525 bp in size (Hiratsuka *et al.* 1989). Over time, chloroplast genomes have experienced dramatic variation, but some conserved structures have been maintained within land plants (Wu and Ge 2012, Tang *et al.* 2010). Many highly conserved genes fundamental to plant life and more variable regions were found on the cp genome, which have been informative over broad time scales (Tong *et al.* 2016). At present, the highly conserved gene order, stable gene content, and slow rate of mutation in chloroplast genomes as an important genetic resource have been used to explore evolutionary variation in land plants (De Las Rivas *et al.* 2002). The chloroplast (cp) genome contains three functional categories which include protein-coding genes, introns and intergenic spacers; the latter two do not encode proteins and are often referred to as non-coding regions but the genes of non-coding regions show that their nucleotide substitution rates are 2.3 times higher than those of protein-coding genes, so some gene sequences from the non-coding regions have been used to study genetic diversity, genetic structures and population structures of Viridiplantae. Thus, the cp genomes of the different species and types of rices are a valuable genetic resource and provide useful information for further studying the genetic variation, population structure, genetic evolution, phylogenetic analysis and breeding programs of Rice (Evenson and Gollin1997).

The present study was aimed to sequence and annotate the whole cp genome of B810S, to reveal the genome structure, gene contents and order of B810S, and to infer phylogenetic relationships of B810S and other rices by using the concatenated amino acid sequences of 99 protein-coding genes.

Materials and Methods

A Rice with pale yellow leaf and belonging to *Oryza* family was collected from a deserted paddy field in Anjiang, Huaihua, Hunan Province, China. This strain Rice was identified as *Oryza sativa* L under a light microscope using the morphology method, then it was named B810S according its collection place. All leaves were cut using a scissor and then washed three times with physiological saline, was stored at -80ºC until use. The total genomic DNA of B810S was isolated from their leaves using a Plant DNA isolation Reagent kit (Tiangen, Beijing, China) following the manual instructions, and were finally stored at -20°C until use. The 1.5% agarose gel with ethidium bromide (EB) was used to detect the integrity and purity of DNA. The DNA sample was quantified via a Qubit Fluorometer (3.0) (Thermo Fisher Scientific, US). DNA sample was disrupted into fragments randomly via the ultrasonic method. End repair, A-tailing, index adapter adding, amplification, and purification were performed for library constriction according to the manufacturer's instructions. All obtained libraries were sent to commercial sequencing via an Illumina HiSeq X sequencing system at Personalbio in Shanghai, China.

To obtain high accurate sequencing clean data, all the obtained raw reads were filtered with quality score (Q <10) (90%), uncalled bases ("*N*" characters) (>10%), and duplicated sequences.

All obtained sequences were assembled into a long sequence using the computer program DNMAN, and was then aligned by the computer program Clustal X 1.81. The cp genome annotation of B810S were performed by comparing with *Oryza sativa* having whole cp genome from *Oryza* genus (Genebank accession number: NC_031333). The computer program MAFFT 7.122 (Katoh and Standley 2013) was used to identify gene boundaries. Translation initiation and mtranslation termination codons were identified based on comparison with those of reported previously. Sequences of protein-coding genes were translated into amino acid sequences using the plant genetic code in MEGA 6.0 (Kumar *et al.* 2008). The secondary structures of the tRNA genes were predicted using Dual Organellar GenoMe Annotator (DOGMA) with default parameters (Wu *et al.* 2017), and the rRNA genes were predicted by comparison with those of

reported genome previously. The circular cp genome of B810S genomic map was drawn via OGDraw v1.2 (Lohse *et al.* 2007).

In order to study the phylogenetic relationships of B810S and other rices from *Oryza* genus, the protein-coding genes sequences were translated into corresponding amino acid sequences, and then they were concatenated. All translated amino acid sequences were aligned using MAFFT 7.122 with default parameters and the ambiguously aligned regions of the amino acid sequences were excluded using Gblocks online server (http://molevol.cmima.csic.es/castresana/Gblocks server. html) with the default parameters. Eight rices from *Oryza* genus from GenBnk in NCBI were selected to perform phylogenetic analyses using the Neighbor-Joining (NJ) method with Kimura two-parameter analysis and bootstrap analysis of 1000 replicates (MEGA 6.0) (Ge *et al.* 1999). The available Rice were *Oryza sativa* (NC_031333), *Oryza longistaminata* (KF359907), *Oryza meridionalis* (KF359906), *Oryza barthii* (KF359904), *Oryza glberrima* (KF359903), *Oryza rufipogon* (KF359902), *Oryza glumipatula* (KR364803) and *Oryza nivara* (KM088022), respectively. *Triticum aestivum* (NC_002762) was used as an outgroup (Table 1). Moreover, The Akaike information criterion (AIC) as implemented in ProtTest 2.4 was used to choose the most suitable model of evolution(Akaike 1974). The resulted phylogenetic tree was drawn using the program FigTree v.1.4 (http://tree.bio.ed.ac.uk/software/figtree).

Table 1. All information of the selected Rice and *Triticum aestivum* using to phylogenetic analysis.

Plant names	Accession number	Authors	Length	Country
Oryza sativa	NC_031333	Kim KH	134502	South Korea
Oryza longistaminata	KF359907	Gao J	134557	China
Oryza meridionalis	KF359906	Gao J	134553	China
Oryza barthii	KF359904	Gao J	134581	China
Oryza glberrima	KF359903	Gao J	134661	China
Oryza rufipogon	KF359902	Gao J	134587	China
Oryza glumipatula	KR364803	Kim KH	134575	South Korea
Oryza nivara	KM088022	Kim KH	134516	South Korea
Triticum aestivum	NC_002762	Ogihara Y	134545	USA

Results and Discussion

The obtained sequences were assembled for a complete cp genome which is a typically circular DNA molecule with 134546 bp in sized (Fig. 1). A total of 148 genes are on the circular cp genome which includes 99 protein-coding genes (PCGs), 41 tRNA genes, 8 rRNA genes and 1 non-coding (control or AT-rich) regions. The assembled, aligned and complete cp genome of B810S has been submitted to the GenBank with the accession number: MW289404. The cp genome was consisted of two strands that are majority strand (H-strand) and minority strand (L-strand). All genes are located on H-strand and L-strand, respectively. Comparing to cp genome of other Rice, the length of the cp genome of B810S was normal, and the gene content and order of cp genome were identical to those of Rice from *Oryza* family. Moreover, nucleotide composition of B810S were 30.6 (A), 30.4 (T), 19.6 (G) and 19.4% (C), respectively; the A+T content is obviously higher than G+C content, and this cp genome is AT skews, which is in accordance with cp genomes of other Rice reported in previous study (Chen *et al.* 2015). According to related documents (Guisinger *el al.* 2011) showed that the length of cp genome depend on the length of

the non-coding regions, many researches thought that the non-coding region involves in gene replication and transcription but not in protein translation.

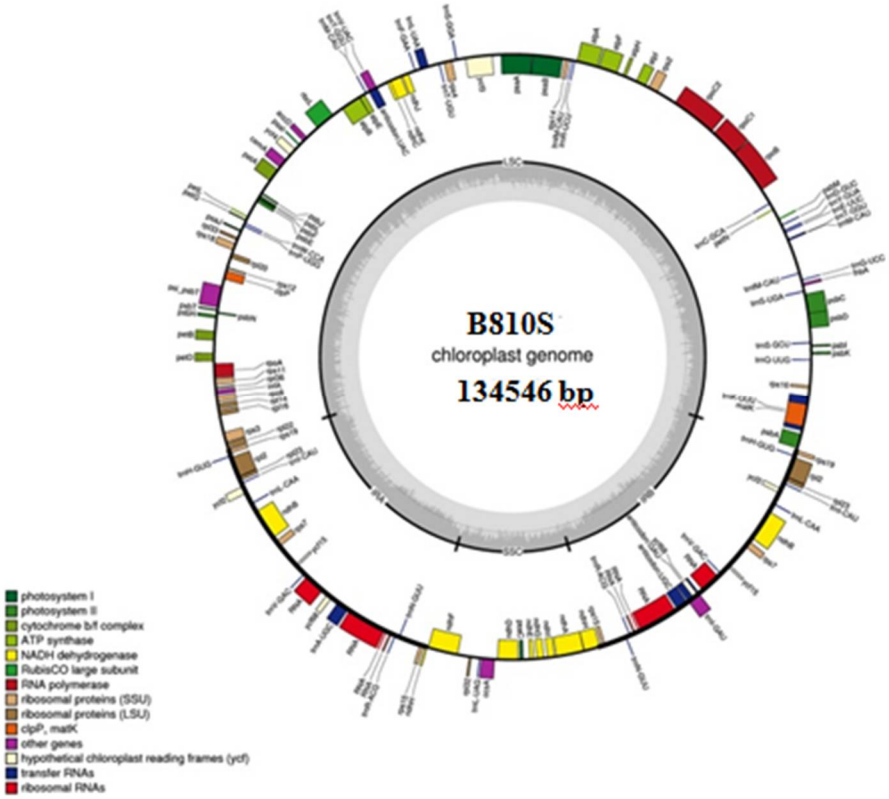

Fig. 1. The complete cp genome of B810S.

Many overlapping regions with different length in this cp genome were found by annotating, which varied from 17 bp to 2204 bp. The largest overlapping region was ndhF gene (2205 bp) locating between ndhH gene and rpL32 gene, but the smallest overlapping region was ORF34 gene (105 bp) that was locate between ORF46 gene and ORF28 gene. Moreover, total of 133 intergenic spacer regions with different length were found between different genes on this genome. The length of intergenic spacer regions ranged from 2 bp to 2583 bp; the shortest intergenic spacer region was location in ndhA and ndhH and the longest intergenic spacer region was location in tRNA-Val and rps7.

The cp genome of B810S has 99 protein-coding genes (PCGs). The length of all PCGs is 62637 bp, accounting for 50.3% of the cp genome; these PCGs encode 20879 amino acids. In the cp genome,45 out of 99 PCGs are located on H-strand, the rest of PCGs are reside on L-strand. The longest PCGs is rpoC2 (4542 bp) which encodes 1514 amino acids, while the shortest PCGs is orf23 (72 bp) that encodes 24 amino acids. Furthermore, All PCGs applied complete codon as their initiator and termination codon. Among all PCGs genes, most of them used the ATG as their initiator codon, most of PCGs used the TAG as their termination codon. No incomplete codons were applied on this cp geome.

A total 8 ribosomal genes with different length were found on the cp genome of B810S. 23S located between tRNA-Ala and 4.5S is large ribosomal gene, and it was 2887 bp in size, and the 4.5S with 94 bp located between 23S and 5S was the smallest ribosomal gene. Additionaly, 41 tRNA genes with different length (57 bp-2759 bp) were found on this cp genome. The longest tRNA gene was tRNA-Lys with 2759 bp, but the shortest was tRNA-Met of 57 bp. All tRNAs had the canonical clover-leaf secondary structure, except tRNA-Met that lacks the dihydrouridine (DHU) arm.

The phylogenetic relationships of B810S and other members of *Oryza* family were conducted by NJ methods with GTR+I+G4 model. Results of the present study showed that B810S and other selected rices from *Oryza* genus were located on one clade of this phylogenetic tree without higher node value support (Fig. 2), indicating that the B810S and other *Oryza* genus rices have close phylogenetic relationships, suggesting that the B810S belongs to *Oryza* genus.

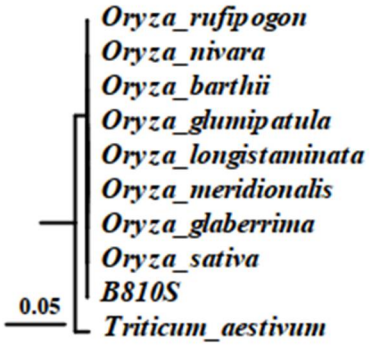

Fig. 2. Phylogenetic relationship of B810S and other *Oryza* genus rices was inferred by NJ analyses using the cp genome, with *Triticum aestivum* as out-group.

The Rice B810S have close phylogenetic relationships with other Rice from *Oryza* family. Moreover, the results of the present study not only can contribute to the molecular identification and taxonomy of Rice, but also provide much useful information and many gene markers for studying the genetic diversity and genetic structure of Rice in different geographical origins in China and other regions of the world in further.

Acknowledgements

This work was supported in part by Scientific Research Fund of Hunan Provincial Education Department (No.2019JJ70046).

References

Asaf Sajjad, Muhammad Waqa, Abdul L. Khan, Muhammad A. Khan,Sang-Mo Kang, Qari M. Imran, Raheem Shahzad, Saqib Bilal, Byung-Wook Yun and In-Jung Lee 2017. The Complete Chloroplast Genomeof Wild Rice (*Oryza minuta*) and Its Comparison to Related Species. Frontiers in Plant Science **8**: 1-15.

Akaike H 1974. A new look at the statistical model identification. IEEE Trans Automatic Control **19**(6): 716-723.

Bao Y, Ge S 2008. Historical retrospect and the perplexity on the studies of the Oryza polyploids. J. Syst. Evol. **46**: 3-12.

Chen J, Hao Z, Xu H, Yang L, Liu G and Sheng Y 2015. The complete chloroplast genome sequence of the relict woody plant *Metasequoia glyptostroboides* Hu et Cheng. Front. Plant Sci. **6**: 447.

De Las Rivas J, Lozano JJ and Ortiz AR 2002. Comparative analysis of chloroplast genomes: functional annotation, genome-based phylogeny, and deduced evolutionary patterns. Genome Res. **12**(4): 567-583.

Daniell H 2007. Transgene containment by maternal inheritance: effective or elusive? Proc Natl. Acad. Sci. U SA **104**: 6879-80.

Evenson RE and Gollin D1997. Genetic resources, international organizations, and improvement in rice varieties. Econ. Dev. Cult. Change **45**: 471-500.

Group II 2012. New grass phylogeny resolves deep evolutionary relationships and discovers C 4 origins. New Phytol. **193**: 304-312.

Ge S, Sang T, Lu BR, Hong DY 1999. Phylogeny of rice genomes with emphasis on origins of allotetraploid species. Proc. Natl. Acad. Sci. U.S.A. **96**: 14400-14405.

Guisinger MM, Kuehl JV, Boore JL and Jansen RK 2011. Extreme reconfiguration of plastid genomes in the angiosperm family geraniaceae: rearrangements, repeats, and codon usage. Mol. Biol. Evol. **28**:1543-1543.

Hagemann R 2010. The foundation of extranuclear inheritance: plastid and mitochondrial genetics. Mol. Genet. Genomics **283**(3): 199-209.

Hiratsuka J, Shimada H, Whittier R, Ishibashi T, Sakamoto M, Mori M, Kondo C, Honji Y, Sun CR and Meng BY 1989. The complete sequence of the rice (*Oryza sativa*) chloroplast genome: intermolecular recombination between distinct tRNA genes accounts for a major plastid DNA inversion during the evolution of the cereals. Mol. Gen. Genet. **217**(2-3): 185-194.

Izawa T 2008. The process of rice domestication: a new model based on recent data. Rice **1**: 127-134.

Jacquemin J, Bhatia D, Singh K and Wing RA 2013. The international oryza map alignment project: development of a genus-wide comparative genomics platform to help solve the 9 billion-people question. Curr. Opin. Plant Biol. **16**:147-156.

Kovach MJ, Sweeney MT and McCouch SR 2007. New insights into the history of rice domestication. Trends Genet **23**: 578-587.

Kohler S, Delwiche CF, Denny PW, Tilney LG, Webster P, Wilson RJ, Palmer JD and Roos DS 1997. A plastid of probable green algal origin in Apicomplexan parasites. Sci. **275**(5305): 1485-1489.

Kato T, Kaneko T, Sato S, Nakamura Y and Tabata S 2000. Complete structure of the chloroplast genome of a legume, Lotus japonicus. DNA Res. **7**(6):323-330.

Katoh K and Standley DM 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. **30**: 772-80.

Kumar S, Nei M, Dudley J, Tamura K 2008. MEGA: a biologistcentric software for evolutionary analysis of DNA and protein sequences. Brief. Bioinformatics **9**: 299-306.

Kim KH, Lee JK, Koh HJ and Yang TJ 2017. High throughput and simultaneous assembly of complete chloroplast and nuclear ribosomal DNA sequences from plant genomes.

Li ZK, Rutger JN 2000. Geographic distribution and multilocus organization of isozyme variation of rice (*Oryza sativa* L.). Theor. Appl. Genet. **101**:379-387.

Lohse M, Drechsel O and Bock R 2007. Organellar Genome DRAW (OGDRAW): a tool for the easy generation of high-quality custom graphical maps of plastid and mitochondrial genomes. Curr. Genet. **52**: 267-274.

Palmer JD, Jansen RK, Michaels HJ, Chase MW and Manhart JR 1988. Chloroplast DNA variation and plant phylogeny. Ann. Mo. Bot. Gard. **75**(4):1180-1206.

Sang T and Ge S 2007. Genetics and phylogenetics of rice domestication. Curr. Opin. Genet. Dev. **17**: 533-538.

Wang S and Gao LZ 2016. Complete chloroplast genome sequence and annotation of the tropical *japonica* group of Asian cultivated rice (*Oryza sativa* L.) Genome Announc **4**(1): e01703-15.

Sugiura M 1989. The chloroplast chromosomes in land plants. Ann. Rev. Cell Biol. **5**: 51-70.

Tong Wei, Kim Tae-Sung and Park Yong-Jin 2016. Rice chloroplast genome variation architecture and phylogenetic dissection in diverse oryza species assessed by whole-genome resequencing. Rice **9**: 57.

Turmel M, Otis C and Lemieux C 1999. The complete chloroplast DNA sequence of the green alga Nephroselmis olivacea: insights into the architecture of ancestral chloroplast genomes. Proc. Natl. Acad. Sci. USA **96**(18): 10248-10253.

Tang L, Zou XH, Achoundong G, Potgieter C, Second G and Zhang DY 2010. Phylogeny and biogeography of the rice tribe (Oryzeae): Evidence from combined analysis of 20 chloroplast fragments. Mol. Phylogenet Evol. **54**: 266-277.

Wu ZQ and Ge S 2012.The phylogeny of the BEP clade in grasses revisited: Evidence from the wholegenome sequences of chloroplasts. Mol. Phylogenet Evol. **62**: 573-578.

Wu ZQ, Gu CH, Tembrock LR, Zhang D and Ge S 2017. Characterization of the whole chloroplast genome of *Chikusichloa mutica* and its comparison with other rice tribe (Oryzeae) species. PLoS ONE **12**(5): e0177553.

(Manuscript received on 6 July, 2021; revised on 6 October, 2021)