Gene 640 (2018) 86–96

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Complete structure and variation of the chloroplast genome of *Agropyron cristatum* (L.) Gaertn

Ning Chen[a,b], Li-Na Sha[a], Zhen-Zhen Dong[a], Chao Tang[a], Yi Wang[a], Hou-Yang Kang[a], Hai-Qin Zhang[a], Xue-Bin Yan[c], Yong-Hong Zhou[a,b], Xing Fan[a,b,*]

[a] Triticeae Research Institute, Sichuan Agricultural University, Chengdu, Sichuan 611130, China
[b] Key Laboratory of Crop Genetic Resources and Improvement, Ministry of Education, Sichuan Agricultural University, Yaan 625014, Sichuan, China
[c] College of Animal Science and Technology, Yangzhou University, Yangzhou 225009, Jiangsu, China

ARTICLE INFO

Keywords:
Agropyron cristatum
Chloroplast genome
Dn/ds ratio
Evolutionary rate
Phylogeny

ABSTRACT

Agropyron cristatum (L.) Gaertner, a perennial grass in the tribe Triticeae (Poaceae), is a wild relative of cereal crops that is suitable for genetic improvement. In this study, we first sequenced the complete chloroplast (cp) genome of *Ag. cristatum* using Hiseq4000 PE150. The *Ag. cristatum* chloroplast genome is 135,554 bp in length, has a typical quadripartite structure and contains 76 protein-coding genes, 29 tRNA genes and four rRNA genes. The cp genome of *Ag. cristatum* was used for comparison with other seven Triticeae species. One large variable region (800 bp), which primarily contained the *rpl*23 (non-reciprocally translocated from IRs) and *acc*D genes, was detected between *rbc*L gene and *psa*I gene within LSC region. The deletion of the *acc*D and translocated *rpl*23 genes in *Ag. cristatum* indicated an independent gene-loss events or an additional divergence in Triticeae. Analyses of the dn/ds ratio and K2-P's genetic distance for 76 protein-coding genes showed that genes with evolutionary divergence might suffer from the effect of sequence regional constraints or gene functional constraints in Triticeae species. Our research will generally contribute to the knowledge of plastid genome evolution in Triticeae.

1. Introduction

The chloroplast (cp) is one of the most vital organs in green plant cells. It has many important functions, such as starch synthesis, nitrogen metabolism, sulfate reduction, fatty acid synthesis, and DNA and RNA synthesis (Zeltz et al., 1993). Chloroplast DNA (cpDNA) is composed of a single circular DNA molecule with a quadripartite structure. In angiosperms, cp genomes range from 115 to 165 kb in length and consist of two copies of inverted repeat (IR) region, a large-single-copy (LSC) region and a small-single-copy (SSC) region (Raubeson and Jansen, 2005; Wicke et al., 2011). Because of the uniparental mode of inheritance and high conservation in gene content and genome structure, the chloroplast genome is generally treated as a single locus (Hirosawa et al., 2004; Raubeson and Jansen, 2005; Nock et al., 2010). With a smaller effective population size and essentially recombination-free, the chloroplast genome has a shorter coalescent time than nuclear genomes (Birky et al., 1983). Also, cp genome sequences have been used for plant phylogenomics and inferring the origin and diversification pattern of species (Parks et al., 2009). Based on cp phylogenomic reconstruction of *Triticum/Aegilops* species in Triticeae, Gornicki et al. (2014) suggested that chloroplast haplotypes were often shared by species or subspecies within major lineages and between the lineages. The chloroplast genome of hexaploid wheat was considered to be contributed by the B genome donor and that this unknown species diverged from *Ae. speltoides* about 980,000 years ago (Middleton et al., 2014). While these studies add to our understanding of cp phylogenomics of annual Triticeae plants, little is known about the cp genomic structure and variation in perennial Triticeae species. With the advent of high-throughput sequencing technologies, it is now inexpensive to obtain cp genome sequences and promote cp-based phylogenomics (Zhang et al., 2016).

Agropyron Gaertn, an important perennial genus in the tribe Triticeae, includes approximately 15–20 species (Kellogg, 2015) and naturally distributes in Europe and Asia. As an excellent wild germplasm for wheat breeding and forage grass (Asay and Johnson, 1990; Limin and Fowler, 1987; Ford-Lloyd et al., 2011; Dong et al., 1992), *Agropyron* was introduced into America during the last century (Clayton et al., 2006; Johnson, 1986). Due to its strong colonization ability,

Abbreviations: PE150, paired end 150 bp long reads; Cp, chloroplast; IR, inverted repeat; LSC, large-single-copy; SSC, small-single-copy; SAUTI, Sichuan Agricultural University; CTAB, Cetyltrimethyl Ammonium Bromide; DOGMA, Dual Organellar GenoMe Annotator; dn, non-synonymous; ds, synonymous

* Corresponding author at: Triticeae Research Institute, Sichuan Agricultural University, Wenjiang 611130, Sichuan, China.
E-mail address: fanxing9988@163.com (X. Fan).

http://dx.doi.org/10.1016/j.gene.2017.10.009
Received 16 June 2017; Received in revised form 27 September 2017; Accepted 5 October 2017
Available online 13 October 2017
0378-1119/

Agropyron cristatum gradually became the dominant invasive species in Northern America (Watson and Dallwitz, 1992). The species of *Agropyron* are usually allopolyploids with the P genome, and chromosome numbers ranged from $2n = 2\times = 14$ to $2n = 6\times = 42$ (Dewey, 1984). The P genome can be combined with the St genome from *Pseudoroegneria* (Nevski) Á. Löve and the Y genome from an unknown donor in various allopolyploid combinations, including the StP genome (*Douglasdeweya*) and the StYP genome (*Kengyilia*) (Yen and Yang, 2011; Jensen, 1990, 1996; Jensen and Chen, 2008; Zhou, 1994; Zhang et al., 1998, 2000; Baum and Johnson, 2008; Fan et al., 2012). Analyses of chloroplast *mat*K, *rbc*L, and *trn*H-*psb*A sequences suggested that *Agropyron* was the most likely maternal genome donor during the speciation of several *Kengyilia* polyploids with the StYP genomes (Zhang et al., 2009; Zeng et al., 2010; Luo et al., 2012a, Sha et al., 2010; Sha et al., 2014;). However, data from *trn*L-F and *rbc*L sequences indicated that species of *Pseudoroegneria* (St-genome donor) served as the maternal donor of some species within *Elymus sensu* lato (including *Kengyilia*) (Liu et al., 2006; Dong et al., 2013). Additional studies on the role of *Agropyron* as a maternal ancestor in the evolution of natural polyploidy are thus needed. Since the P genome of *Agropyron* is a basic component of several polyploid genera in Triticeae, more attention should be paid to the different processes of polyploidization.

In this study, we performed the cp genome sequencing of *Ag. cristatum* using Hiseq4000 PE150. Knowledge of the cp genomic variation and phylogeny of *Ag. cristatum* will provide a better understanding of the evolution of natural polyploid individuals. The objectives of this study were (1) to detect variations in the cp genome of *Ag. cristatum*; and (2) to assess the evolutionary rate of cp genes among Triticeae species;

2. Materials and methods

2.1. DNA sources/data collection

Agropyron cristatum collected from Kazakhstan (Accession No. PI 598628) was used in this study. The seed material was kindly provided by the American National Plant Germplasm System (Pullmam, Washington, USA). The plants and voucher specimens are deposited at the Herbarium of the Triticeae Research Institute, Sichuan Agricultural University, China (SAUTI). The *Triticum aestivum* (accession number NC002762) chloroplast sequence (Ogihara et al., 2000, 2002) downloaded from NCBI was used as a reference sequence for the assembly of *Ag. cristatum*. Comparisons included seven published Triticeae genomes in NCBI (*Aegilops speltoides*, KJ614405; *Aegilops tauschii*, KJ614412; *Hordeum jubatum*, KM974741; *Hordeum vulgare*, EF115541; *Secale cereale*, KC912691; *Triticum monococcum*, LC005977; *Triticum urartu*, KJ614411) and *Bromus vulgaris* (KM974737).

2.2. DNA extraction and sequencing

Total genomic DNA of *Ag. cristatum* was extracted from the actively growing fresh leaves using a modified CTAB (Cetyltrimethyl Ammonium Bromide) method (Doyle, 1990). Ground leaves were placed in CTAB buffer with 2% polyvinylpyrrolidone at 65 °C for 40 min. The solution was centrifuged, and the supernatant was mixed with an equal volume of chloroform/isoamyl alcohol mixture (24:1) and centrifuged twice. DNA was precipitated with isopropanol at − 20 °C for an hour, and then pelleted by centrifugation, washed twice with 70% ethanol, aspirated and re-suspended in double-distilled water. The solution was then treated with RNase (Fermentas) at 37 °C for 1.5 h and DNA extracted using the same procedure as above. DNA quantity and RNA contamination were verified with ethidium bromide on 1% agarose.

The PCR product was fragmented for construction of short insert (500 bp) libraries according to the NEBNext®. Total genomic DNA was sequenced using HiSeq 4000 PE150. Approximately 63,676,400–600 bp paired-end reads were generated from the sequencing library.

2.3. Genome assembly and annotation

De novo assembly was performed with identified chloroplast-related reads using SPADES v3.6.1 (Bankevich et al., 2012), CLC Genomics Workbench 8 and SOAPdenovo2 (Luo et al., 2012b). The putative plastid contigs were identified using the mapping algorithm and assembled in Geneious 8.1 (Kearse et al., 2012). The genome was annotated using Dual Organellar GenoMe Annotator (DOGMA) (Wyman et al., 2004). A circular cp genome map of *Ag. cristatum* was drawn with the Organellar Genome DRAW (Lohse et al., 2013).

2.4. Comparison of nine cp genomes

The software mVista was used to compare the cp genome of *Ag. cristatum* with eight other genomes (*Ae. speltoides*, *Ae. tauschii*, *H. vulgare*, *H. jubatum*, *S. cereale*, *T. monococcum*, *T. urartu*, *B. vulgar*) using the annotation of *Ag. cristatum* as a reference (Frazer et al., 2004). The 9 cp genome sequences were aligned with MAFFT v6.833 (Katoh and Toh, 2010) using default settings.

2.5. Identification of repetitive sequences

The modiWed script, SSRhunter (Zhu et al., 2003), was used to search for simple sequence repeats (SSRs) ranging from mono-, to penta- nucleotide repeats. A web-based Tandem Repeats Finder (http://tandem.bu.edu/trf/trf.html) was used to analyze the tandem repeats with minimal lengths of 28 bp, sequence identity > 90%, and edit distances of < 3 bp (Benson, 1999). The tandem repeats finder used parameter settings of 2 for matches and 7 for mismatches and indels. The minimum alignment score and maximum period size were set at 50 and 500, respectively. All the identified repeats were manually verified and nested or redundant results were removed.

2.6. Analysis of dn/ds and K2-P's genetic distance

To analyze non-synonymous (dn) and synonymous (ds) substitution rates, the same individual functional protein-coding exons were extracted and aligned separately by using DNAMAN (version5.2.10, Lynnon Biosoft). The non-synonymous (dn) and synonymous (ds) substitution rates for each protein-coding exon were estimated in MEGA 6.0 (Tamura et al., 2013). All coding genes were compared among the eight species (except for the outgroup *B. vulgaris*), and we calculated the singleton variable sites and the parsimony informative sites with the DnaSP5.1 software (Rozas et al., 2003).

The sequences of 76 protein coding genes were analyzed and aligned by DNAMAN. The interspecies genetic distance was calculated based on the Kimura 2-parameter model using MEGA 6.0.

2.7. Phylogenetic analysis

Phylogenetic analyses were performed for the eight Triticeae species using maximum likelihood (ML) and Bayesian inference (BI). *Bromus vulgaris* was used as the outgroup. Two analyses were carried out based on 1) the complete cp DNA sequences and 2) concatenated protein-coding sequences. A total of 76 coding genes from nine taxa were compiled into a single file (concatenated protein-coding sequences) and aligned with MAFFT.

ML analysis was performed using PAUP*4.0b10 (Swofford, Sinauer Associates, http://www.sinauer.com). The evolutionary model used for the three different data matrixes was determined using ModelTest v3.0 with Akaike information criterion (AIC) (Posada and Crandall, 1998). The optimal models identified were TVM + G + I for the complete cp DNA sequence data and GTR + I for the protein-coding sequence data. ML heuristic searches were performed with 100 random addition sequence replications and TBR branch swapping algorithm. The robustness of the trees was estimated by bootstrap support (BS) (Felsenstein,