Limited Polymorphisms between Two Whole Plastid Genomes in the Genus *Zizania* (Zizaniinae)

Zhiqiang Wu1*, Cuihua Gu1*, Luke R Tembrock1 and Song Ge1

1 State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences, Beijing, China
2 School of Landscape and Architecture, Zhejiang Agriculture and Forestry University, Hangzhou 311300, P.R. China.
3 Department of Biology, Colorado State University, Fort Collins, Colorado, USA

* Contributed equally to this work

Abstract

Plastids, originated from cyanobacteria through endosymbiosis, contain their own DNA genome (plastome), and are uniparentally inherited in most of plant species. In this study, we reported the complete plastome sequence of Manchurian wild rice (*Zizania latifolia*) obtained by traditional Sanger sequencing and compared it with the previously published plastome of North America wild rice (*Z. aquatica*) from the same genus. The plastome of *Z. latifolia* has a total sequence length of 136,461 bp exhibiting a typical circular structure including a pair of 20,878 bp inverted repeats (IRa, b) separated by a large single-copy region (LSC) of 62,115 bp and a small single-copy region (SSC) of 12,590 bp, and it is only 97 bp longer than North America wild rice. The gene content, order, and orientation are similar to all other grass species. Four junctions in two plastome structures are exactly the same between the *Zizania* species. From complete genome comparisons, 744 (568, 46 and 130 for LSC, IR and SSC) substitutions are found between the two *Zizania* plastomes including 267 from coding regions and 477 from non-coding regions. Insertions/deletions (Indels) are mainly found in non-coding regions (138) except the only one repeat indel from coding regions in *rpoC2* gene. The most informative biomarkers (*rps16-fmQ, ndhF* and *matK*) are also the most divergent between the two *Zizania* species. The completed plastome and comparative analyses from these two *Zizania* species would be as important resource for further systematic or population genetic studies in the Zizaniinae subtribe, marker assisted breeding, and plastomic transformation.

Keywords: *Oryzeae; Zizania; PCR; Plastome; Polymorphisms; Indels

Introduction

The tribe Oryzeae contains some of the most economically important species in the world. The tribe is divided into two well supported subtribes named Oryzinae and Zizaniinae [1-4]. Economically important species are found in both subtribes such as cultivated rice in Asian and African (*Oryza sativa* and *O. glaberrima*) from Oryzinae and the wild rice (*Zizania latifolia* and *Z. aquatica*) in Asia and North America from Zizaniinae [5,6]. Besides rice (*Oryza sativa*), many other close relatives in this tribe could provide valuable genetic resources for rice breeding and improvement. For example, *Z. latifolia* had been used for generating introgression lines for potential breeding applications [7,8]. Furthermore, with the development of high throughput sequencing technologies, more species in the *Oryza* genus have been completed their whole genomes [9-13]. These progresses have made the tribe Oryzeae as an ideal system to study genomic evolution [14-16].

Now, in Oryzinae, numerous plastid genomes have been completely sequenced [17,18], however, the finished plastomes from subtribe of Zizaniinae are limited. *Zizania* is a small genus only comprised of four species including *Z. palustris*, *Z. aquatica*, and *Z. texana* native to North America, and one species *Z. latifolia* widely distributed in Eastern Asia [19]. As the only species native to Asia, *Z. latifolia*, named Manchurian wild rice, is used as a food plant with both the stem and grain being consumed in several Asian countries. As a relative of rice, *Zizania* could be an important genetic resource in rice breeding and genetic transformation [7]. Plastid transformation can accelerate genetic modifications, and therefore, the finished plastomes are particularly important genetic resource in developing improved rice varieties. Plastomes have been shown to be a very effective tool in genomic transformation with fewer concerns regarding gene flow into wild or non-transformed individuals [20].

Plastomes, similar to animal mitochondrial genomes, maintain a conserved circular double-stranded DNA structure, with sizes ranging from 115 to 165 kb in land plants [21,22], and stable gene content and order [23]. These features of plastomes and the decreasing cost of high-throughput sequencing technologies have led to an increase in the number of completed plastomes. At present, more than 900 species' completed plastomes are currently available in the NCBI database (http://www.ncbi.nlm.nih.gov/genomes/). Besides being an effective tool for genetic transformation [24,25], complete plastomes are also important resources to explore the genetic and evolutionary variation between plant groups. For example, as a primary source for plant molecular systematic and taxonomic studies, plastomes have provided strong phylogenetic signals at multiple levels of inquiry [17,26-28]. For instance, the markers from plastomes have been used to explore the biogeographical relationships among plant populations [27,29] and for DNA barcoding [24,25].

In this study, we combined the traditional polymerase chain reaction (PCR) and Sanger sequencing methods to fully sequence and assemble the whole plastome of *Z. latifolia*. We conducted comprehensive comparisons with *Z. aquatica* [30] and other rice species to infer what evolutionary changes have occurred within the Oryzeae tribe. The finished plastome of *Z. latifolia* will provide important genetic information for plastid transformation involved with crop improvement.

*Corresponding author: Zhiqiang Wu, Department of Biology, Colorado State University, Fort Collins, Colorado, USA, E-mail: wu.zhiqiang.1020@gmail.com
Cuihua Gu, School of Landscape and Architecture, Zhejiang Agriculture and Forestry University, Hangzhou 311300, P.R. China, E-mail: gu_cuihua@126.com

Received October 26, 2015; Accepted November 13, 2015; Published November 17, 2015


Copyright: © 2015 Wu Z, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.
Methods and Materials

Complete plastid genome sequence of *Zizania latifolia*

Fresh leaves of *Zizania latifolia* were collected from plants grown in the greenhouse of the Institute of Botany of the Chinese Academy of Sciences in Beijing. Total cellular DNA was extracted using the cetyltrimethyl ammonium bromide (CTAB) method [31] and purified with phenol extraction to remove proteins. With PCR amplification and Sanger sequencing methods, we completely sequenced the plastid genome (plastome) of *Z. latifolia*. For sequencing *Z. latifolia*, a full set of primers was designed based on two previously sequenced bamboo plastomes [17,32]. PCR amplification and purification of the products were performed, as described in Tang et al. [3]. The purified products were directly sequenced on an ABI 3730 automated sequencer (Applied Biosystems, Foster City, CA, USA). The sequences were assembled with the ContigExpress program from the Vector NTI Suite 6.0 (Informax Inc., North Bethesda, MD).

Plastid genome annotation and drawing

The fully assembled plastid genome of *Z. latifolia* was annotated using the DOGMA (Dual Organellar GenoMe Annotator, [33]). The first draft annotation from the DOGMA output was subsequently manually inspected and adjusted for accurate assessment of the start and stop codons and the exon–intron boundaries of genes. Both tRNA and rRNA genes were identified by BLASTN searches against the database of chloroplast genomes from NCBI. The final annotations were plotted as a circular genome by the bioinformatics tools Circos 0.67 [34] (Figure 1). For comparison, we downloaded the published plastid genomes from five species in Oryzeae and one bamboo species [17,35].

Dynamic variation of junction regions in plastid genome

Based on the configuration of the plastid genome with the two inverted repeat regions (IRa and IRb) and the two single copy regions (LSC and SSC) [36,37], four junctions named as JLa, JLb, JSa, and JSb are located between the two single copy (LSC and SSC) regions and the two IRs (IRa and IRb) [38]. The dynamic variation at the four junction regions (JLa, JLb, JSa, and JSb) can contribute to the size variation of plastid genomes [39,40]. The detailed IR border positions and the distance with adjacent genes among seven grass species plastomes (*Oryza sativa* ssp. *japonica* [AY522330], *O. australiensis* [KJ830774], *Leersia tisserantii* [JN415112], *Z. latifolia* [in this study], *Z. aquatica* [KJ870999], *Rhynchoryza subulata* [JN415114], *Phyllostachys propinqua* [JN415113]) were therefore compared in this work.

![Figure 1: Chloroplast genome information and variation maps of *Z. latifolia* (KT161856). From outside to inside, tracks depict: 1) coding genes on forward strand; 2) coding genes on reverse strand; 3) the number and distribution of substitutions (simply named as SNP) (grey bar color); 4) the number and distribution of non-repeat insertion/deletions (Indels) (green bar color); 5) the number and distribution of poly structures (grey bar color); 6) the number and distribution of repeat Indels (green bar color). The coding genes are colored based on functional group with different color codes at the bottom. Maps were generated with software Circos v0.67.](image-url)
Polymorphism analysis from coding and non-coding regions

To comprehensively compare the complete plastomes of two Zizania species, we employed the mVISTA program under the Shuffle-LAGAN mode [41] to detect whole genome variation with the plastome of O. sativa ssp. japonica (AY522330) as the reference. To identify the polymorphic regions that may be highly informative for phylogenetic or population genetic analyses from the two Zizania plastid genomes, we divided the whole plastid genome as coding and non-coding regions. Excluding the non-variation of the RNA genes, all protein coding genes were separately analyzed as coding regions, and all intron and intergenic regions were examined as non-coding regions (only the sequence longer than 200 bp were used in this analysis). These regions were extracted by using custom PERL scripts from the whole plastid genome and aligned using ClustalW under MEGA6 [42] and adjusted manually using the similarity criterion [43] for non-coding regions or preserving the reading frames for coding regions. The aligned sequences were used to calculate the sequence identity values (SI) by using BioEdit software [44] and the number of the polymorphic sites were reported from DnaSP v5.10 [45]. Only one part of the IRs regions was used in this process given the identical nature of the other repeat.

Phylogenetic analysis

To determine the phylogenetic relationships among the six Oryzeae species and outgroup bamboo species, as whole plastomes alignment was used to build phylogenetic tree. Based on the collinearity of the whole plastid is extremely conserved among grass family [17,46], the whole plastomes of seven species were aligned in MAFFT v7.221 [47] and phylogenetic analysis was implemented in PAUP* 4.0b10 [48], Bayesian inference methods were used to infer relationships: maximum parsimony (SI) by using BioEdit software [44] and the number of the polymorphic sites were reported from DnaSP v5.10 [45]. Only one part of the IRs regions was used in this process given the identical nature of the other repeat.

Table 1: Comparison of major features of seven Poaceae plastome.

<table>
<thead>
<tr>
<th>Subfamily</th>
<th>Species</th>
<th>Total size</th>
<th>LSC region</th>
<th>IR region</th>
<th>SSC region</th>
<th>Gene Content</th>
<th>GenBank Accession</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Length (bp)</td>
<td>GC (%)</td>
<td>Length (bp)</td>
<td>GC (%)</td>
<td>Length (bp)</td>
<td>GC (%)</td>
</tr>
<tr>
<td>Ehrhartioideae</td>
<td>Oryza sativa ssp. japonica</td>
<td>134,551</td>
<td>39.00</td>
<td>80,604</td>
<td>37.11</td>
<td>20,802</td>
<td>44.35</td>
</tr>
<tr>
<td></td>
<td>Oryza australisensis</td>
<td>134,549</td>
<td>39.98</td>
<td>80,614</td>
<td>37.07</td>
<td>20,796</td>
<td>44.36</td>
</tr>
<tr>
<td></td>
<td>Leersia tisserantii</td>
<td>136,550</td>
<td>38.88</td>
<td>81,865</td>
<td>37.01</td>
<td>21,329</td>
<td>44.05</td>
</tr>
<tr>
<td></td>
<td>Zizania latifolia</td>
<td>136,461</td>
<td>39.00</td>
<td>82,115</td>
<td>37.13</td>
<td>20,878</td>
<td>44.42</td>
</tr>
<tr>
<td></td>
<td>Zizania aquatica</td>
<td>136,364</td>
<td>39.02</td>
<td>82,013</td>
<td>37.14</td>
<td>20,879</td>
<td>44.41</td>
</tr>
<tr>
<td></td>
<td>Rynchospora subulata</td>
<td>136,303</td>
<td>39.00</td>
<td>82,029</td>
<td>37.14</td>
<td>20,840</td>
<td>44.36</td>
</tr>
</tbody>
</table>
| Bambusoideae     | Phyllostachys^
|                  | 139,704    | 38.88      | 83,227    | 36.96      | 21,800    | 44.23           | 127          |

* Sequenced in this study.

Phylogenetic analysis

To determine the phylogenetic relationships among the six Oryzeae species and outgroup bamboo species, as whole plastomes alignment was used to build phylogenetic tree. Based on the collinearity of the whole plastid is extremely conserved among grass family [17,46], the whole plastomes of seven species were aligned in MAFFT v7.221 [47] and phylogenetic analysis was implemented in PAUP* 4.0b10 [48], Bayesian inference methods were used to infer relationships: maximum parsimony (SI) by using BioEdit software [44] and the number of the polymorphic sites were reported from DnaSP v5.10 [45]. Only one part of the IRs regions was used in this process given the identical nature of the other repeat.

Phylogenetic analysis

To determine the phylogenetic relationships among the six Oryzeae species and outgroup bamboo species, as whole plastomes alignment was used to build phylogenetic tree. Based on the collinearity of the whole plastid is extremely conserved among grass family [17,46], the whole plastomes of seven species were aligned in MAFFT v7.221 [47] and phylogenetic analysis was implemented in PAUP* 4.0b10 [48], Bayesian inference methods were used to infer relationships: maximum parsimony (SI) by using BioEdit software [44] and the number of the polymorphic sites were reported from DnaSP v5.10 [45]. Only one part of the IRs regions was used in this process given the identical nature of the other repeat.

Polymorphisms between the two Zizania plastomes

Discovery of polymorphisms between plastomes are the basis of marker selection in plant systematic or barcoding research [24,52], in addition, discovery of sites with limited variation could be the candidates of plastid transformation [20]. We conducted the following genome wide and local specific analysis to fully investigate the polymorphisms variation between two Zizania plastomes.

First, based on the conserved features of the land plant plastomes [22], the polymorphisms were discovered from the whole genome alignment of two Zizania plastomes and categorized as either

Table 2: The number of polymorphisms in different regions by comparison from two Zizania plastome.

<table>
<thead>
<tr>
<th>Type</th>
<th>Region</th>
<th>Coding Regions</th>
<th>Non-Coding Regions</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNP</td>
<td></td>
<td>LSC</td>
<td>IR</td>
<td>SSC</td>
</tr>
<tr>
<td></td>
<td></td>
<td>189</td>
<td>8</td>
<td>70</td>
</tr>
<tr>
<td></td>
<td></td>
<td>379</td>
<td>38</td>
<td>60</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>744</td>
</tr>
<tr>
<td>Indel</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LSC</td>
<td></td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>63</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>32</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>124</td>
</tr>
<tr>
<td>IR</td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>6</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>SSC</td>
<td></td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>7</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>72</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>34</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>137</td>
</tr>
</tbody>
</table>

Sequencing, assembly and annotation

Compared with the low-cost and high-throughput next generation sequencing (NGS) technologies, PCR and Sanger sequencing produced highly creditable results with fewer errors [37]. Meanwhile, some other features made plastid are easy to amplify including that the plastome is small but with abundant component in cellular DNA, and rates of nucleotide substitution are relatively slow [50]. By integrating the overlapped Sanger sequenced fragments (longer 100 bp) from PCR products, we successfully assembled the whole plastid genome for Z. latifolia with 136,461 bp in length by using the method from Wu et al., [32] (Figure 1). After the automatic annotation from DOGMA [33], comparative analyses, and manual verification, the annotated plastid genome of Z. latifolia was uploaded to GenBank with accession KT161956 (Figure 1). It is composed of two single-copy regions separated by a pair of inverted repeats (IR) of 20,878 bp each, which account for 30.60% of the whole plastid genome. The large single copy (LSC) and the small single copy (SSC) regions span 82,115 bp and 12,590 bp, respectively. The proportion of LSC and SSC length in the total plastid genome is 60.17% and 9.23% respectively, and those features were all extremely similar with the other six published plastomes (Table 1). The same gene content, gene number and gene order among the seven species reflects the highly conserved plastomes of Poaceae [46,51]. The size differences among plastomes are mainly a result of the variation within intergenic sequences.
substitutions (for simply named as SNPs) or insertions/deletions (Indels, Table 2). From the results (Table S1 and S2), we found that the large single copy (LSC) regions possessed the most variable sites (568 SNPs and 124 Indels), and the two inverted repeats (IR) contained the least variable sites (46 SNPs and 6 Indels). However, when we consider the number of variable sites per kilo base pairs (kbp), the small single copy (SSC) region was the highest with 10.3 substitutions/kbp, and 6.9 and 2.2 for the LSC and IR. This result was the same as the evolutionary rate variation among the three regions [53]. When comparing polymorphisms between coding and non-coding regions, the non-coding regions contained the most variable sites in the LSC and IR regions, but in SSC, the number of SNPs was higher in coding regions. This result shows that in LSC and SSC regions, the rate of variation between coding and non-coding is different.

Second, for comparison the whole genome variation within the two Zizania plastomes, we also employed the mVISTA program

![Figure 2: Identity plot](image-url)

*Figure 2:* Identity plot that compares the chloroplast genomes of the two Zizania data sets used in this study with *O. sativa* ssp. *Japonica* (AY522330) as the reference sequence. The vertical scale indicates the percentage of identity to the reference, ranging from 50% to 100%. The horizontal axis indicates the coordinated base position within the chloroplast genome. Genome regions are color coded as protein-coding, RNA (rRNA and tRNA), intron, and conserved noncoding regions.
[41]. We used *O. sativa ssp. japonica* (AY522330) as the reference for these comparisons (Figure 2). In regards to structural differences, the organization of the plastome was conserved between the two *Zizania* species with no differing translocations or inversions found. The two IR regions were more conserved than the LSC and SSC regions with 99.8% similarity found between them, and the coding regions were more conserved than non-coding regions (by the view from different colors). The most variation from coding regions was found in the *rpoC2* gene (Figure 3). After the alignment of the amino acid (AA) sequence of *rpoC2*, the variable regions were focused around 650 to 820 AA. The length of this gene was different in all seven species surveyed and was mainly caused by size variation of repeat sequence within *rpoC2*. For the two *Zizania* species, there was one only 7 AA deletion around position 680 found in *Z. aquatica*. Among the non-coding regions (the pink portions in Figure 2), *ndhC-trnV*, *rps16-trnQ* and *ycf3-trnS* demonstrated the most divergent patterns between the two *Zizania* species.

Third, to examine the sequence identity (SI value) of the two *Zizania* plastomes, the whole plastome was divided into individual coding (gene regions) and non-coding (intergenic and intronic regions). Among 76 annotated protein coding genes (Table S3), which were the same in gene content as other grass species [17], the SI values for the two *Zizania* species were all larger than 98% with the average value for SI being 99.5% with 23 genes showing no variation. These results showed that all coding genes were extremely conserved between two *Zizania* species. However, the genes *matK*, *ndhF* and *rpoC2* possessed the lowest SI values, and these regions have been used as universal markers in barcoding and phylogenetic studies [54], which also show phylogenetic signal at lower taxonomic levels. For the SI values from 126 non-coding regions (only one copy of IR regions was used) (Table S4), we found that the average SI values across all non-coding regions longer than 200 bp was 97.9% and was only a little lower than coding regions. Only three regions (*ndhC-trnV*, *rps16-trnQ* and *ycf3-trnS*) had SI values lower than 90%. *Rps16-trnQ* had the greatest number of polymorphic sites and as such was also an important marker for plant systematic studies [54].

**Conserved variation of two *Zizania* plastome junction boundaries**

Based on the quadripartite structure of plastomes from most land plants [22] (Figure 4A), four junctions (JLA, JLB, JSA, and JSB) were named between the two IRs (IRa and IRb) and the two single copy (LSC and SSC) regions (Figure 4C) [38]. It has already been shown that the dynamic variation of IR regions can contribute to the size variation between plastomes [22] and as such provide useful phylogenetic signals [38]. By combining the phylogenetic relationships among the seven grass species plastomes species (Figure 4B) and applying the distance between border position and the adjacent genes, we were able to compare the gene distance in a phylogenetic context (Figure 4C). By comparison with the variation from the *Oryza* genus [37], the two species of *Zizania* retained the same distance around the four junctions (Figure 3B). This indicates that the distance between gene and junction border is more conserved in *Zizania* than *Oryza*. We also examined the difference at gene distance to junction between the two subtribes. The distance of JSA and JSB reflects the variation of the two Oryzeae subtribes from this plastome feature. In subtribe Zizaniinae, the two genes distance from JSA is 319 bp (*rps15*) and 89-93 bp (*ndhF*). However, in subtribe Oryzinae, those distances were 301 bp and 41-42 bp (except for *L. tisserantii*). This feature could be used as a marker to separate the two subtribes. In total, the junction in JLA and JLB are more conserved than JSA and JSB.

**Figure 3:** Partial alignment of amino acid from *rpoC2* gene indicating the length variation caused by the repeat sequences in seven grass species.
Acknowledgement

This work was supported by the National Natural Science Foundation of China (30990240). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Reference


