Genome analysis

Simple sequence repeats in organellar genomes of rice: frequency and distribution in genic and intergenic regions

Passoupathy Rajendrakumar, Akshaya Kumar Biswal, Sena M. Balachandran, Kommoju Srinivasarao and Raman M. Sundaram*

Biotechnology Laboratory, Crop Improvement Section, Directorate of Rice Research, Rajendranagar, Hyderabad-500030, India

Received on April 15, 2006; revised on September 23, 2006; accepted on October 21, 2006
Advance Access publication October 31, 2006

ABSTRACT

Motivation: Simple sequence repeats (SSRs) are abundant across genomes. However, the significance of SSRs in organellar genomes of rice has not been completely understood. The availability of organellar genome sequences allows us to understand the organization of SSRs in their genic and intergenic regions.

Results: We have analyzed SSRs in mitochondrial and chloroplast genomes of rice. We identified 2528 SSRs in the mitochondrial genome and average 870 SSRs in the chloroplast genomes. About 8.7% of the mitochondrial and 27.5% of the chloroplast SSRs were observed in the genic region. Dinucleotides were the most abundant repeats in genic and intergenic regions of the mitochondrial genome while mononucleotides were predominant in the chloroplast genomes. The rps and nad gene clusters of mitochondria had the maximum repeats, while the rpo and ndh gene clusters of chloroplast had the maximum repeats. We identified SSRs in both organellar genomes and validated in different cultivars and species.

Contact: rms_28@rediffmail.com.

Supplementary information: Supplementary Data are available at Bioinformatics online.

1 INTRODUCTION

Simple sequence repeats (SSRs) are ubiquitous, hypervariable and abundant in many prokaryotic and eukaryotic genomes. Polymorphism in SSRs is generally believed to be the result of replication error (Moxon and Wills, 1999), which occurs at a rate higher than the mutation in non-repetitive DNA (Wierdl et al., 1997). The expansion of repeats in the coding region of the gene or their presence in regulatory regions may alter the gene expression (Cummins and Zoghbi, 2000). The 5'-untranslated region (5'-UTR) of the waxy gene in rice contains a poly(CT) motif whose length polymorphism is associated with amylose content (Ayers et al., 1997). The property of maternal inheritance of the mitochondria and its low nucleotide substitution rate (Wolfe et al., 1987) were effectively used to elucidate genetic relationships among Hordeum species (Nishikawa et al., 2002). A recent report describes phylogenetic analysis based on mononucleotide repeats and flanking nucleotide sequences from the organellar genomes (Nishikawa et al., 2005). Availability of genome sequences in the public domain has removed technical and economic limitations and dramatically accelerated the process of development of SSR markers. The DNA sequence information of nuclear and organellar genomes of rice has led to efficient and high-throughput in silico identification of SSR loci. Yet, very little is known on the nature and organization of SSRs in genic and intergenic regions of organellar genomes. We have therefore, analyzed SSRs in organellar genomes of rice for their occurrence, nature, organization and distribution in both coding and non-coding regions. We have also designed primers for a few SSRs, to assess their polymorphic potential in a set of rice genotypes including different cultivars and species. These markers can be utilized further in phylogenetic studies and synteny analysis among cereals.

2 METHODS

The DNA sequences from GenBank (http://www.ncbi.nlm.nih.gov) were used for the purpose of generation of SSR data. These sequences include the mitochondrial genome of Oryza sativa subsp. japonica (cv. Nipponbare—GI# 47118326) and chloroplast genomes of O.sativa subsp. indica (cv. 93-11—GI# 42795473), O.sativa subsp. japonica (cv. Nipponbare—GI# 42795537) and Oryza nivara, a wild species of rice (GI# 50233947). SSRs were identified and localized using the software SSR Identification Tool (SSRIT) (Temnykh et al., 2001), which identifies perfect di-, tri-, tetra-, penta- and hexanucleotide repeats. We have considered only those repeats, wherein the motif was repeated more than 3 times for further analysis. Mononucleotide repeats (with a repeat length of ≥6 nt) were identified using the software FastPCR (Kalendar, 2006, http://www.biocenter.helsinki.fi/blt/programs/fastpcr.htm). The rationale for choosing a low cut-off value is that SSRs are often disrupted by single base substitution (Subramanian et al., 2003). The occurrence of repeats in genic and intergenic regions was identified based on the sequence annotation information available in GenBank database. To compare the observed number of SSRs with the expected number, we calculated the expected number of SSRs using the formula given by de Wachter (1981):

\[ N(M) = p(M)[1 - p(M)]\{(N[1 - p(M)] + 2L) \}

\[ N' = N - mL - 2L + 1, \]

where, \( M \) is the repeat unit (repeat type), \( N(M) \) is the expected number of times that, in a DNA segment of length \( N \), we find \( t \) consecutive \( Ms \). \( L \) is the length of \( M \) and \( p(M) \) is the probability of \( M \) (obtained by multiplying the probability of each nucleotide contained within the repeat unit \( M \)).

To test the utility of organellar SSRs as potential genetic markers, a few SSRs with a maximum repeat length of ≥12 nt were selected. Primers were designed using the software FastPCR with standard parameters as described.

*To whom correspondence should be addressed.
by Chen et al., 1997. Homology search for matches in the nuclear genome of rice was performed by using the BLAST algorithm (Altschul et al., 1997, http://www.ncbi.nlm.nih.gov/blast) and only those primer pairs that did not show any match were considered for further analysis. These primer pairs (35 and 5 for mitochondria and chloroplast, respectively) were used for amplification of DNA isolated from three indica accessions (Abhaya, GEB 24 and Madhukar), three japonica accessions (Azucena, Nipponbare and Taipei 309), two basmati accessions (Taraari Basmati and Pusa Basmati-1) and one accession each of Oryza nivara, Oryza rupifogon, Oryza longistaminata and Porteresia coarctata.

Total DNA was isolated from leaves of 18–20 day-old rice plants following the standard protocol (Kochert et al., 1989). PCR was carried out in 10 μl reaction volume containing 25 ng of template DNA, 0.2 μM of each primer, 200 μM of each dNTP, 10 mM Tris–HCl (pH 8.3), 50 mM KCl, 1.5 mM MgCl2 and 1 U of Taq polymerase in a 96-well thermal cycler (Eppendorf, Germany). The thermal profile was: 94°C for 5 min, followed by 35 cycles of 94°C for 30 s, 55°C for 30 s, 72°C for 1 min and a final extension at 72°C for 7 min. Amplified products were resolved in 6% denaturing PAGE followed by silver staining (Panaud et al., 1996). The visualized SSR alleles were precisely sized using the software utility Alphaseq® (Alphainnotech, USA) with 50 bp ladder (MBI Fermentas, Lithuania) as the size standard. Polymorphism information content (PIC) values were calculated by using the online resource, PIC Calculator (http://www.agri.huji.ac.il/~weller/Hayim/parent/PIC.htm).

### 3 RESULTS AND DISCUSSION

#### 3.1 SSRs in the mitochondrial genome

Approximately 3.5% of the mitochondrial genome of japonica harbored SSRs. Total of 2528 repeat motifs were identified, of which 220 (8.7%) were localized in genic regions (Table 1). The density of SSRs was 28 bp/kb in genic regions, while it was 36 bp/kb in intergenic regions. On average, for every kb of DNA, the genic region had 4.2 SSRs and the intergenic region had 5.3 SSRs (Supplementary Figs 1 and 2). Of the 104 genes/coding sequences with known function (Notsu et al., 2002), 58 genes (~56%) possessed SSRs. Among the classes of repeats, dinucleotides were the most abundant (~48% of repeats), with frequent occurrence of the repeat motif AG/CT.

Mononucleotides (~40% of repeats) were the second most abundant repeats. Among the mononucleotide repeats, poly(A) or (T) was far more abundant than poly(C) or (G). Trinucleotides (272) were the next most copious repeats with frequent occurrence of the repeat motif AAG. In addition, 42 tetra- and 10 penta- and a single hexanucleotide repeat motif were also detected. The proportion of different classes of repeats found in the mitochondrial genome is shown in Supplementary Figure 3. Only 95 out of 1200 dinucleotide repeats and 18 out of 272 trinucleotide repeats were present in the genic region. The AT/TA motif was the most abundant dinucleotide repeats present in the genic region, followed by TC/GA (Supplementary Fig. 4). Total of 18 trinucleotide repeats were also present in the genic region of the mitochondrial genome. Polymorphism in trinucleotide repeats has been shown to have a direct correlation with genetic disorders and other important functions in the human genome (Subramanian et al., 2003). In addition to di- and trinucleotide repeats, the genic region also possessed two tetra- and a single hexa- repeat. Among the important gene clusters, the maximum numbers of SSRs were found in rps (30) and nad (28) gene clusters (Supplementary Table 1).

Cytoplasmic male sterility (CMS) has been suggested to be the result of disruption of cellular respiration machinery since the production of male gamete is an energy intensive process (Dai et al., 1978; Chen and Liang, 1991). The mitochondrial genome is known to code for many genes associated with energy metabolism (Race et al., 1999) and any modifications in the mitochondrial DNA may be associated with the trait of CMS (Mignouna et al., 1987). Our study identified the presence of many SSRs in genes associated with cellular respiration and polymorphism in some of them could be associated with male sterility.

### 3.2 SSRs in the chloroplast genomes

On average, 4.5% of the chloroplast genomes was found to be composed of SSRs. The chloroplast genomes of indica, japonica and O.nivara possessed 869, 875 and 871 SSRs, respectively. Chloroplast genome of indica had the least SSRs (235) in the genic region while that of O.nivara had the most (342) (Table 1). The density of SSRs was 35–37 bp/kb in the genic region and 47–51 bp/kb in the intergenic region. For every kilobase pair of DNA, 5.3 SSRs were observed in the genic region and 7.0–7.6 SSRs in the intergenic region. (Supplementary Figs 1 and 2). So far, total of 68, 70 and 92 genes/coding sequences have been characterized in the chloroplast of indica (Tang et al., 2004), japonica (Hiratsuka et al., 1989) and O.nivara (Masood et al., 2004), respectively. About 67% of the indica chloroplast genes possessed SSRs while it varied from 60% in japonica to 69% in O.nivara. The maximum numbers of SSRs were found in rpo and ndh gene clusters (45–79) (Supplementary Tables 2–4). Mononucleotides were the most abundant repeats (~63%) with frequent poly(A) or (T). Dinucleotides were the next predominant repeats (~30%) in the chloroplast genomes (269 in japonica, 268 in indica and 267 in O.nivara). The AT/TA was the most abundant dinucleotide repeat motif. Each of three chloroplast genomes possessed 42 trinucleotide repeats; TTC being predominant. Both

<table>
<thead>
<tr>
<th>Repeat motif</th>
<th>Mitochondria Japonica</th>
<th>Chloroplast Japonica</th>
<th>O.nivara</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mono</td>
<td>104</td>
<td>145</td>
<td>160</td>
</tr>
<tr>
<td>Genic</td>
<td>899</td>
<td>403</td>
<td>394</td>
</tr>
<tr>
<td>Intergenic</td>
<td>1105</td>
<td>196</td>
<td>179</td>
</tr>
<tr>
<td>Di</td>
<td>95</td>
<td>72</td>
<td>90</td>
</tr>
<tr>
<td>Genic</td>
<td>1105</td>
<td>196</td>
<td>179</td>
</tr>
<tr>
<td>Intergenic</td>
<td>254</td>
<td>26</td>
<td>20</td>
</tr>
<tr>
<td>Tri</td>
<td>18</td>
<td>16</td>
<td>22</td>
</tr>
<tr>
<td>Genic</td>
<td>40</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>Intergenic</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Tetra</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Genic</td>
<td>40</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>Penta</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Genic</td>
<td>10</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Intergenic</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Hexa</td>
<td>1</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Genic</td>
<td>—</td>
<td>1</td>
<td>—</td>
</tr>
<tr>
<td>Intergenic</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Total</td>
<td>2528</td>
<td>869</td>
<td>875</td>
</tr>
<tr>
<td>Frequency per kb</td>
<td>5.2</td>
<td>6.5</td>
<td>6.5</td>
</tr>
<tr>
<td>Density (bp/kb)</td>
<td>35.6</td>
<td>43.6</td>
<td>43.9</td>
</tr>
</tbody>
</table>
japonica and *O.nivara* had 9 tetranucleotide repeats while indica had 10 such repeats. Interestingly, a single pentanucleotide repeat was observed in both indica and *O.nivara* genomes. A single hexanucleotide repeat was found unique to japonica. The proportion of different classes of repeats in chloroplast genomes is shown in Supplementary Figure 3.

The distribution of SSRs in genic regions showed some interesting patterns. The AT/TA motif was the most predominant dinucleotide repeat in the genic region (Supplementary Fig. 5) similar to liverworts, maize and pea chloroplasts (Powell et al., 1996). While *O.nivara* possessed significantly more dinucleotides repeats (103) in the genic region, japonica (90) and indica (73) had less repeats in the genic region. A relatively higher number of genes reported in *O.nivara* as compared to indica and japonica might be the reason for more dinucleotides repeats. *O.nivara* had a significantly higher number of trinucleotide repeats (26) in the genic region compared to indica (16) and japonica (22). Both indica and *O.nivara* had two identical tetranucleotide repeats in the genic region. In contrast, japonica possessed only a single tetranucleotide repeat in the genic region that was similar to the one of the two tetranucleotide repeats present in other two genomes. A single pentanucleotide repeat (aaagt), was present in the genic region of indica and *O.nivara* while japonica possessed a unique hexanucleotide repeat (atagaa). These observations strongly support the theory of origin of cultivated rice proposed by Second (1985) who hypothesized that indica might have originated from *O.nivara*. It will be desirable to develop markers based on these unique motifs to enable easy identification of the maternal origin of rice accessions, particularly in these two subspecies.

### 3.3 Comparison of SSRs in mitochondrial and chloroplast genomes

Based on the length of repeat motif, SSRs are generally classified into hypervariable markers (class I, repeat length of \( \geq 20 \) nt), potentially variable markers (class II, repeat length of 13–20 nt) and stochastic markers (class III, repeat length of 6–12 nt) (McCouch et al., 2001). Of the total number of SSRs identified in this study, 85 (japonica mitochondria), 18 (indica chloroplast), 15 (japonica chloroplast) and 16 (nivara chloroplast) SSRs may be grouped in to class II type markers (Supplementary Table 5). The absence of Class I SSRs in these genomes may be due to their smaller size, a relatively stable nature and low mutation rates. Among the mononucleotide repeats, poly(A) or poly(T) were far more abundant than poly(C) or poly(G) in both organellar genomes. These findings are consistent with previous observations about the difference in abundance of mononucleotide repeats (Temnykh et al., 2001; Subramanian et al., 2003). Dinucleotide repeats were predominant in the mitochondrial genome while mononucleotide repeats were more frequent in the chloroplast genomes. Further analysis on SSRs (di, tri, tetra etc.) in genic and intergenic regions revealed some differences between mitochondria and chloroplast genomes. In the genic region of mitochondria, dinucleotide repeats were higher in proportion than others. In chloroplast genomes, mononucleotides outnumbered all other repeats. In contrary, trinucleotides were reported to be abundant in the genic region of the nuclear genome of rice. The differential pattern of dinucleotides in the genic region may be due to the fact that these two genomes use independent DNA polymerase machinery and different methods of replication (Campbell et al., 1999). However, abundance of dinucleotides was also reported in the ESTs of some dicotyledonous species (Kumpatla and Mukhopadhyay, 2005).

Among the dinucleotide repeats, the CG/GC repeats were extremely rare in genic and intergenic regions of the organellar genomes. This observation was similar to that of nuclear SSRs of rice (McCouch et al., 2002) and organellar SSRs of other organisms (Wills et al., 2005). The AT/TA motif was the most abundant in the genic region of mitochondria and chloroplasts. SSRs with AT repeats have been reported as the most abundant and highly polymorphic class of repeats in rice nuclear genome also (McCouch et al., 2002). Among the trinucleotide repeats, the motif AAG was the most common in mitochondria while chloroplast genomes had comparatively higher frequency of TTC. Mitochondria possessed 53 different types of trinucleotide repeats whereas the chloroplast had only 17. In mitochondria, 24 different types of tetranucleotide repeats were observed while only 9 different types were observed in chloroplast. The limited number of repeat types in each category may be once again attributed to a relatively smaller size and stable nature of organellar genomes. The hexanucleotide repeat, observed in mitochondria (aaataa), was found to be different from the one observed in chloroplasts (atagaa). Only the mitochondrial hexanucleotide repeat was present in the genic region. In both organellar genomes, dinucleotides were repeated up to 6 times while tri- and tetra-nucleotides were repeated 4 times each. The penta- and hexanucleotide repeats were observed three times each. The implications of excess of short iterated repeats (<8 repeat units) could be extremely important for genomic stability and for the evolution of additional genomic features, such as codon usage (Field and Wills, 1998).

The expected number and frequency of major classes of SSRs (mono-, di- and tri-) were calculated by de Wachter formula. The observed frequencies of SSRs were much lower than the expected frequencies in intergenic regions of mitochondrial and chloroplast genomes (Supplementary Table 6). In contrast, the genic regions had more abundant trinucleotide SSRs than the expected. As the translation of genes relies on triplet codons, it would have led to higher observed trinucleotide SSRs.

The potential of a few SSRs were analyzed through PCR in a set of rice accessions as described in Section 2. All SSR markers showed clear and robust amplification. Thirteen mitochondrial SSRs and one chloroplast SSR exhibited polymorphism (Supplementary Table 7 and Supplementary Fig. 6). The PIC value ranged from 0.14 to 0.45 for the mitochondrial SSRs while the chloroplast SSR showed a value of 0.34. The observation of lower polymorphism in chloroplast SSRs is in concurrence with the earlier observation by Ishii et al. (2001) on a low PIC value.

### 4 CONCLUSION

Our study has helped to identify genes possessing SSRs in organellar genomes. The repeat motifs are not uniformly distributed across the genomes but mostly confined to intergenic regions. Hence, the occurrence of SSRs is a non-random event. Most of the mitochondrial genes associated with respiration have also been found to possess SSRs, which could be targeted for the development of PCR-based markers to study the maternally inherited traits like CMS. Similarly, a few of the genic SSRs have been found to be significantly different between indica and japonica chloroplast
genomes, which may be helpful in developing indica or japonica specific PCR markers. The SSR markers developed in the present study could also be useful in determining the maternal origin of rice accessions and in phylogenetic studies.

ACKNOWLEDGEMENTS

The authors sincerely thank Dr B.C. Viraktamath, Project Director, Directorate of Rice Research, Hyderabad for the facilities and encouragement provided for carrying out the study. The authors also thank Dr K. Muralidharan and Dr C.N. Neeraja for critically reviewing the manuscript.

Conflict of Interest: none declared.

REFERENCES


