Genome-wide intraspecific DNA-sequence variations in rice
Bin Han* and Yongbiao Xue†

Genome-wide comparative analysis of the DNA sequences of two major cultivated rice subspecies, Oryza sativa L. ssp indica and Oryza sativa L. ssp japonica, have revealed their extensive microcolinearity in gene order and content. However, deviations from colinearity are frequent owing to insertions or deletions. Intraspecific sequence polymorphisms commonly occur in both coding and non-coding regions. These variations often affect gene structures and may contribute to intraspecific phenotypic adaptations.

Address

*National Centre for Gene Research, Shanghai Institutes of Biological Sciences, Chinese Academy of Sciences, 500 Caobao Road, Shanghai 200233, China
e-mail: bhan@ncgr.ac.cn
†Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Datun Road, Andingmenwai, Beijing 100101, China
e-mail: ybxue@genetics.ac.cn

This review comes from a themed issue on Genome studies and molecular genetics
Edited by Takaji Sasaki and Ronald R Sederoff
1369-5266/03/$ – see front matter © 2003 Elsevier Science Ltd. All rights reserved.
DOI 10.1016/S1369-5266(03)00004-9

Abbreviations
BAC bacterial artificial chromosome
EST expressed sequence tag
GLA4 Guangluai 4
Indels insertions/deletions
IRGSP International Rice Genome Sequencing Project
SNP single nucleotide polymorphisms

Introduction
The past two years have been a time of harvest for rice genome research. In April of 2000, Monsanto announced that it had produced a draft sequence of the rice genome, and that it would share the data with individual researchers and the International Rice Genome Sequencing Project (IRGSP) consortium [1]. Moreover, a comprehensive rice transcript map, which includes a large number of mapped expressed sequence tags (ESTs), and an integrated physical and genetic map for rice have been constructed [2,3]. These advances have greatly contributed towards the construction of a minimal tilling path of the large-inserted bacterial artificial chromosome (BAC) or P1-derived artificial chromosome (PAC) clones that are being used to completely sequence the rice genome.

Domesticated rice (Oryza sativa L.) cultivars belong to two major subspecies indica and japonica. Cultivated indica varieties have dispersed throughout the tropics and subtropics from Eastern India, whereas japonica varieties moved northward from Southern China and developed into temperate ecotypes [4]. Indica and japonica separated more than 1 million years ago [5] and are partially isolated in terms of sexual reproduction. Intraspecific phenotypic variations, including growth, developmental and environmental adaptations, are apparent between these two cultivated subspecies [6]. Molecular dissection of these variations will not only increase our understanding of phenotypic adaptations but also facilitate the genetic improvement of cultivated rice and other crops.

As part of the efforts towards these goals, the IRGSP aimed to produce a high-quality genome sequence of the japonica variety ‘Nipponbare’ by adapting a clone-by-clone strategy [7]. A completed phase-II and high-quality draft of the japonica Nipponbare genome was announced on 18th December, 2002. At the time of writing this review, two rice chromosomes have been completely sequenced by the Japanese and Chinese groups [8,9]. Meanwhile, two phase-I draft sequences have recently been reported for the O. sativa L. japonica Nipponbare and indica 9311 cultivars [10,11]. This progress represents a major advance in our understanding of the content and complexity of rice genome [12]. It has revealed several important types of DNA-sequence polymorphism between rice subspecies at the genome-wide level and some important insights into the molecular basis of intraspecific phenotypic variations.

Genome-wide comparative genetic and physical mapping
Various molecular studies have consistently shown a distinct difference between indica and japonica in the quantification of genomic DNA and repetitive sequence [13]. Nevertheless, extensive genomic colinearity has also been demonstrated between rice subspecies by early genetic and recent physical mapping. A high-density genetic linkage map for rice has been constructed with 2275 markers using a single F2 indica × japonica population [14]. An indica genetic linkage map was also constructed using an indica × indica recombinant inbred population [15]. A set of restriction fragment length polymorphism (RFLP) markers that were used in the construction of both maps allow comparisons between the indica and the indica × japonica linkage maps. Such comparisons reveal that the gene order is highly conserved between the two maps and that both maps have nearly the same total length with regard to the coverage by common
markers. Nevertheless, four small inversions in chromosome arms 1S, 1L, 4L and 8L have been detected [15].

More-refined structural conservation has been identified by comparative physical mapping. Recently, a comparative approach was used in parallel with chromosomal sequencing to construct a fine physical map of chromosome 4 [16**]. A fine BAC-based physical map of japonica Nipponbare chromosome 4 was constructed by integrating 114 sequenced BAC clones from an indica variety ‘Guangluai 4’ (GLA4) with fingerprinted data from the Nipponbare genome [16**]. In this study, the indica sequence scaffolds were used to anchor japonica Nipponbare BACs onto the corresponding regions of chromosome 4 through in silico BAC-end-sequence searches. Although repetitive DNA sequences caused some complications, most of the japonica Nipponbare BACs were rapidly and accurately located on rice chromosome 4 in accordance with the previously anchored indica GLA4 BACs. The indica and japonica physical maps showed an overall synteny. This result clearly demonstrates that physical maps of closely related subspecies can be constructed with reasonable confidence using this comparative approach.

**Intraspecific DNA-sequence polymorphisms**

Recent progress in the genome sequencing of rice subspecies has provided an opportunity to detect genome-wide intraspecific variations in DNA sequence. GLA4, which was widely grown in China during the 1970-80s, is believed to be a typical indica variety. Thus, comparison of the indica GLA4 and japonica Nipponbare genomes is likely reveal some details of intraspecific variations in rice DNA sequence.

Analysis of rice chromosome 4 identified two classes of differences between the Nipponbare and GLA4 sequences: insertion/deletions (Indels) and single nucleotide polymorphisms (SNPs) [9**]. Feng et al. [9**] aligned the sequences of 2.3 Mb of three contiguous segments of indica GLA4 chromosome 4 and 2.4 Mb of its collinear sequences from japonica Nipponbare. The alignments showed extensive sequence co-linearity, but deviations from collinearity frequently occur due to Indels (Figure 1). Within this 2.3-Mb aligned region, the indica sequence had 63 Indels with a total length of 198 kb and the japonica sequence 138 Indels of 312 kb. Among these Indels, only 43% (85 of 198) for indica and 37% (114 of 312) for japonica were co-located with simple sequence repeats that could be identified by the RepeatMask program (Table 1). Many Indels contained entire active genes that were not related to transposons. The insertions usually occur in the intergenic regions, but some of the coding sequences also are interrupted by insertion sequences.

Some polymorphic regions also occur in both the GLA4 and Nipponbare sequences. Sequence polymorphisms between the indica and japonica subspecies are often present in both coding and non-coding regions. A detailed analysis identified 9056 SNPs in the aligned 2.3-Mb GLA4 sequence and 2.4-Mb Nipponbare sequence, indicating an average frequency of one SNP per 256 bp for GLA4 and one SNP per 268 bp for Nipponbare. In total, 2132 SNPs were detected in Nipponbare exons in the homologous region, whereas just 1495 SNPs were found in the GLA4 exons. About 1000 more SNPs were located in the intergenic regions of indica GLA4 than in the same regions of the japonica Nipponbare sequence. Sequence polymorphisms in coding regions that differ between the two subspecies might influence the expression of genes, although further functional tests are required to verify their roles.

**Microcolinearity and its exceptions at gene level**

In the same 2.3-Mb homologous region of chromosome 4, 388 genes were predicted for the indica and 415 genes for
Six peroxidase genes have also been identified in the *japonica* sequence, but they were not detected in the *indica* sequence at the collinear region (B Han, unpublished data). Similarly, a gene encoding a possible gibberellin-induced receptor-like protein kinase was identified in the *indica* but not the *japonica* sequence. The differences in the gene contents of the two genomes may result in the phenotypic variations that are often associated with the two subspecies. By contrast, one pseudogene that is identical to a rice phospholipid-hydroperoxide glutathione peroxidase gene (which encodes 166 amino acids) is found in the collinear regions of both sequences; both pseudogenes are identical and predicted to encode a disrupted 111-amino-acid protein. This finding may reflect the overall conservation of the two genomes.

It appears that *japonica* chromosome 4 is likely larger than that of *indica* because of the expansion of the *japonica* sequence by insertions. If this were also true for the other rice chromosomes, then the *indica* genome will be smaller than that of *japonica* (B Han, unpublished data). Although the *indica* and *japonica* subspecies are very closely related, *japonica* might have evolved earlier than *indica* as more insertions have been added to the chromosomes of *japonica*. Since *indica* and *japonica* rice share a polyphyletic origin in *O. nivara* populations [4], some of the differences identified between the two chromosomes may simply result from differences between the two varieties used. More genomic data from other *indica* and *japonica* varieties are necessary to determine the evolutionary relationships among the cultivated rice species. Comparative information about the chromosome organization of the two closely related rice subspecies has important implications for the development of new molecular markers for genetic mapping. Similar comparative analyses of *Arabidopsis* accessions have shown that both the relocation of genes and sequence polymorphisms between accessions (in both coding and non-coding regions) are common in the *Arabidopsis* genome [17,18]. Intraspecific violation of genetic colinearity has also been identified in maize [19].

Recently, a working draft of the *indica* variety 93-11 genome with 4.3-X coverage redundancy was completed by whole-genome shotgun (WGS) sequencing, allowing an assessment of the degree of colinearity between the *indica* varieties. Feng et al. [9**] identified 7423 contigs with a length of 26 Mb that were located on chromosome 4 from a total of 127 550 contigs in the *indica* 93-11 draft sequence. A super-contig of the draft sequences of chromosome 4 of 93-11 was therefore constructed unambiguously. These contigs matched 7551 contigs on chromosome 4 of *japonica* Nipponbare and covered 75.42% of this chromosome [9**]. They also indicated 128 possible insertions on chromosome 4 [9**]. Similarly, Sasaki et al. [8*] compared a 493-kb contiguous sequence of Nipponbare chromosome 1 with the *indica* 93-11 draft-sequences [8*]. The *indica* 93-11 draft sequence covered 78% of the

---

**Table 1**

<table>
<thead>
<tr>
<th></th>
<th>Indica (cv. Guatingiao)</th>
<th>Japonica (cv. Nipponbare)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of 2.3-Mb homologous regions</td>
<td>388</td>
<td>415</td>
</tr>
<tr>
<td>Genes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gene density (bp per gene)</td>
<td>5979</td>
<td>5846</td>
</tr>
<tr>
<td>Average exons per gene</td>
<td>5.2</td>
<td>5.1</td>
</tr>
<tr>
<td>Average introns per gene</td>
<td>4.2</td>
<td>4.1</td>
</tr>
<tr>
<td>Average exon size (bp)</td>
<td>281</td>
<td>306</td>
</tr>
<tr>
<td>Average intron size (bp)</td>
<td>320</td>
<td>316</td>
</tr>
<tr>
<td>SNPs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total number</td>
<td>9056</td>
<td>9056</td>
</tr>
<tr>
<td>SNPs in exon</td>
<td>1495</td>
<td>2132</td>
</tr>
<tr>
<td>SNP proportion in exon (bp per SNP)</td>
<td>379</td>
<td>304</td>
</tr>
<tr>
<td>SNPs in intron</td>
<td>1655</td>
<td>1,974</td>
</tr>
<tr>
<td>SNP proportion in intron (bp per SNP)</td>
<td>315</td>
<td>272</td>
</tr>
<tr>
<td>SNPs in intergenic region</td>
<td>5906</td>
<td>4950</td>
</tr>
<tr>
<td>Indels</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total number</td>
<td>63</td>
<td>138</td>
</tr>
<tr>
<td>Length of Indels (kb)</td>
<td>198</td>
<td>312</td>
</tr>
<tr>
<td>Repeats length in Indels (kb)</td>
<td>85</td>
<td>116</td>
</tr>
<tr>
<td>Percentage of repeats in whole indels</td>
<td>43%</td>
<td>37%</td>
</tr>
<tr>
<td>Indels &gt;10 kb</td>
<td>3</td>
<td>11</td>
</tr>
<tr>
<td>Indels 1–10 kb</td>
<td>37</td>
<td>25</td>
</tr>
<tr>
<td>Indels &lt;1 kb</td>
<td>23</td>
<td>102</td>
</tr>
<tr>
<td>Total length (bp)</td>
<td>2 319 728</td>
<td>2 426 015</td>
</tr>
</tbody>
</table>
493-kb region with a total of 65 gaps detected in the aligned contigs. Many of the differences between the sequences of Nipponbare chromosomes 1 and 4 and the indica 93-11 draft sequences may be due to the errors and artifacts that are intrinsic to the WGS strategy, but it is likely that some true sequence differences between the two subspecies have been identified. The 4X draft sequence is likely to include some errors, but nucleotide differences between the two indica varieties will eventually be identified when a high-quality sequence of indica 93-11 is available. Therefore, with a due caution, it is feasible that it will be possible to use the completed Nipponbare sequences as a reference genome sequence to analyze and anchor the draft sequences of a closely related species. In addition, some indica and japonica BAC sequences have also been used to identify the mosaic organization of orthologous sequences in grass genomes [20–23].

Conclusions and perspectives
The genome-wide comparative information about the chromosome organization and sequence polymorphism of two closely related rice subspecies has important implications for the development of new molecular markers for genetic mapping. By systematically searching these differences, a complete set of markers that is based on SNPs or other sequence variations could be developed. This will no doubt facilitate map-based gene cloning and help researchers to find allelic variations that are associated with intraspecific phenotypic adaptations in rice cultivars. These intraspecific phenotypic adaptations will impact on molecular rice breeding. In addition, intraspecific sequence comparison can provide a powerful tool for gene annotation and the identification of domestication dynamics.

Rapid progress in the analysis of genome sequences has allowed the systematic identification of intraspecific DNA variations in several plant species including rice. Comparing the genomes of *O. sativa* L. ssp indica and *O. sativa* L. ssp japonica has allowed us to identify differences in genome microstructure. Apart from developing new molecular markers for rice molecular breeding, detailed structural and functional analyses of sequence variations between the two major rice subspecies may eventually lead to a molecular understanding of intraspecific variations in phenotype and adaptations. We believe that fully contiguous and high-quality rice indica and japonica genome sequences are necessary for these endeavors.

Acknowledgements
We thank Y Zhang, Q Zhao, Q Feng and YJ Zhang for their helpful comments and for allowing us to describe some of their unpublished work in this manuscript.

References and recommended reading
Papers of particular interest, published within the annual period of review, have been highlighted as:

- of special interest
- of outstanding interest

6. Morishima H: Evolution and domestication of rice. In Rice Genetics IV. Edited by Khush GS, Brar DS, Hardy B. New Delhi: Science Publishers, Inc.; 2001:63-77. The authors describe some of the implications of rice evolution from the viewpoints of genetics and ecology. They conclude that domestication might be a process driven by the conscious and unconscious selection of adaptive gene blocks that are distributed over the genome.
13. Ohmido N, Kijima K, Akiyama Y, de Jong JH, Fukui K: Quantification of total genomic DNA and selected repetitive sequences reveals concurrent changes in different DNA


This is the first report to describe comparative physical mapping of the indica and japonica genomes. On the basis of the sequence scaffolds of 114 BAC clones of indica Guangluai 4, 29 fingerprinted BAC contigs from japonica Nipponbare were rapidly and accurately anchored on chromosome 4.


