CURRENT OPINION
www.current-opinion.com

# 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.

**Addresses**
*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

**Current Opinion in Plant Biology** 2003, **6**:134–138

This review comes from a themed issue on
Genome studies and molecular genetics
Edited by Takuji Sasaki and Ronald R Sederoff

1369-5266/03/$ – see front matter


**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 $F_2$ *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.

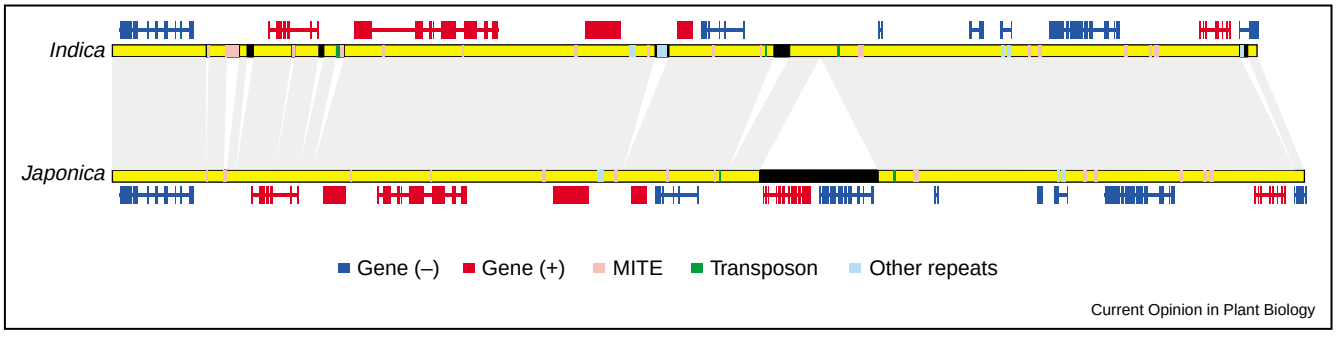
Analysis of rice chromosome 4 identified two classes of differences between the Nipponbare and GLA4 sequences: insertion/deletions (Indels) and single nucleotide polymorphisms (SNPs) [9••]. Feng *et al.* [9••] aligned the sequences of 2.3 Mb of three contiguous segments of *indica* GLA4 chromosome 4 and 2.4 Mb of its collinear sequences from *japonica* Nipponbare. The alignments showed extensive sequence co-linearity, but deviations from colinearity 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

**Figure 1**




Sequence comparison of an orthologous region (of about 100 Kb) from the two cultivated rice subspecies, *Oryza sativa* L. ssp *indica* (cv. GLA4) and *Oryza sativa* L. ssp *japonica* (cv. Nipponbare). Light-gray shading indicates their homologous regions and black bars show the insertions or deletions (Indels) that have occurred in the two subspecies. Repetitive elements are shown by bars of different colors. Predicted gene orders and structures in the top (Gene +) and bottom (Gene –) strands are indicated in dark blue and red, respectively. MITE, miniature inverted-repeat transposable element.

**Table 1**

**Comparison of 2.3-Mb homologous regions between the two rice subspecies.**

| | | Indica (cv. Guangluai 4) | Japonica (cv. Nipponbare) |
|---|---|---|---|
| | Total number of 2.3-Mb homologous regions | 388 | 415 |
| **Genes** | Gene density (bp pergene) | 5979 | 5846 |
| | Average exons per gene | 5.2 | 5.1 |
| | Average introns per gene | 4.2 | 4.1 |
| | Average exon size (bp) | 281 | 306 |
| | Average intron size (bp) | 320 | 316 |
| **SNPs** | Total number | 9056 | 9056 |
| | SNPs in exon | 1495 | 2132 |
| | SNP proportion in exon (bp per SNP) | 379 | 304 |
| | SNPs in intron | 1655 | 1,974 |
| | SNP proportion in intron (bp per SNP) | 315 | 272 |
| | SNPs in intergenic region | 5906 | 4950 |
| **Indels** | Total number | 63 | 138 |
| | Length of Indels (kb) | 198 | 312 |
| | Repeats length in Indels (kb) | 85 | 116 |
| | Percentage of repeats in whole indels | 43% | 37% |
| | Indels >10 kb | 3 | 11 |
| | Indels 1–10 kb | 37 | 25 |
| | Indels <1 kb | 23 | 102 |
| | **Total length (bp)** | 2 319 728 | 2 426 015 |

the *japonica* sequence. The average numbers of exons and introns per gene are almost the same in the *indica* and *japonica* sequences. Furthermore, the gene contents and orders are highly conserved in the two subspecies; however, the gene density is probably higher in *japonica* than in *indica*. The structures of predicted genes in the chromosome 4 sequences of both *indica* and *japonica* are often affected by Indels and SNPs. SNPs were found in the exons and introns of *indica* GLA4 at frequencies of one SNP per 379 and one SNP per 315 bp, respectively. These frequencies were one SNP per 304 bp and one SNP per 272 bp for *japonica* exons and introns, respectively. The results are summarized in Table 1.

Most of the 388 predicted genes in the *indica* chromosome 4 sequence are conserved in the *japonica* sequence, and the additional genes in the *japonica* sequence are derived from retroelements or are hypothetical genes. Most of the hypothetical genes are presumably non-functional as only one or two exons were identified for each of these genes (B Han, unpublished data). Six peroxidase genes have also been found in a cluster in the *japonica* sequence, whereas just a single peroxidase gene is present in the *indica* sequence. Some genes that are unique to either the *indica* or the *japonica* sequence were also found in the collinear region of chromosome 4. A gene known to encode the rice chlorophyl1-binding protein of photosystem II and a gene encoding a serine/threonine protein kinase were 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

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 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

1. Barry G: **The use of the Monsanto draft rice genome sequence in research**. *Plant Physiol* 2001, **125**:1164-1165.

2. •• Wu J, Maehara T, Shimokawa T, Yamamoto S, Harada C, Takazaki Y, Ono N, Mukai Y, Koike K, Yazaki J *et al.*: **A comprehensive rice transcript map containing 6591 expressed sequence tag sites**. *Plant Cell* 2002, **14**:525-535.
Building on the basis provided by previously mapped yeast artificial chromosome (YAC) clones and more than 20 000 sequences of ESTs, the authors map 6591 ESTs onto the rice chromosomes, thereby producing a comprehensive rice transcript map by polymerase chain reaction (PCR) screening. This map presents a detailed view of the arrangements of transcribed genes along the length of the physical map. It has greatly helped to confirm of the location on the chromosomes of several large-inserted BAC or P1-derived artificial chromosome (PAC) clones.

3. •• Chen M, Presting G, Barbazuk WB, Goicoechea JL, Blackmon B, Fang G, Kim H, Frisch D, Yu Y, Sun S *et al.*: **An integrated physical and genetic map of the rice genome**. *Plant Cell* 2002, **14**:537-545.
The authors of this paper describe a highly comprehensive BAC-based physical map that was created by integrating BAC-fingerprinted contigs with unique genetic markers. This work provides a basis for the tilling of BAC clones for Nipponbare genome sequencing.

4. Khush GS: **Origin, dispersal, cultivation and variation of rice**. *Plant Mol Biol* 1997, **35**:25-34.

5. Bennetzen JL: **Comparative sequence analysis of plant nuclear genomes: microcolinearity and its many exceptions**. *Plant Cell* 2000, **12**:1021-1029.

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.

7. Sasaki T, Burr B: **International rice genome sequencing project: the effort to completely sequence the rice genome**. *Curr Opin Plant Biol* 2000, **3**:138-141.

8. • Sasaki T, Matsumoto T, Yamamoto K, Sakata K, Baba T, Katayose Y, Wu J, Niimura Y, Cheng Z, Nagamura Y *et al.*: **The genome sequence and structure of rice chromosome 1**. *Nature* 2002, **420**:312-316.
This paper presents a complete sequence of chromosome 1 of rice *japonica* Nipponbare. The authors describe the annotation of the sequence and the functional classification of the predicted genes. Comparisons between the sequence of a segment of chromosome 1 and the draft sequence of *indica* 93-11 are also discussed.

9. •• Feng Q, Zhang Y, Hao P, Wang S, Fu G, Huang Y, Li Y, Zhu J, Liu Y, Hu X *et al.*: **Sequence and analysis of rice chromosome 4**. *Nature* 2002, **420**:316-320.
These authors present a complete sequence and analysis of chromosome 4 of rice *O. sativa japonica* Nipponbare. They also describe sequence comparisons between *indica* and *japonica* using a contiguous sequence segment of 2.3 megabases of chromosome 4 from a typical *indica* variety, Guangluai 4, and its collinear 2.4-megabase region of the Nipponbare chromosome 4. Intraspecific sequence polymorphisms and Indels between the two subspecies are reported. A detailed analysis of sequence alignment is also described.

10. • Goff SA, Ricke D, Lan TH, Presting G, Wang R, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H *et al.*: **A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*)**. *Science* 2002, **296**:92-100.
The authors describe a 6X draft sequence for the *japonica* Nipponbare variety. They also provide a detailed annotation of the draft sequence.

11. • Yu J, Hu S, Wang J, Wong GK, Li S, Liu B, Deng Y, Dai L, Zhou Y, Zhang X *et al.*: **A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*)**. *Science* 2002, **296**:79-92.
A report on a 4X draft sequence of the genome of the rice variety *O. sativa indica* 9311. This is the first whole-genome draft sequence for an *indica* variety.

12. Bancroft I: **Insights into cereal genomes from two draft genome sequences of rice**. *Genome Biol* 2002, **3**:1015.1-1015.3.

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**

**families in *indica* and *japonica* rice**. *Mol Gen Genet* 2000, **263**:388-394.

14. Harushima Y, Yano M, Shomura A, Sato M, Shimano T, Kuboki Y, Yamamoto T, Lin SY, Antonio BA, Parco A *et al.*: **A high-density rice genetic linkage map with 2275 markers using a single $F_2$ population**. *Genetics* 1998, **148**:479-494.

15. Li WM, Tang DZ, Wu WR, Lu HR, Worland AJ: **A molecular map based on an *indica*/*indica* recombinant inbred population and its comparison with an existing map derived from an *indica*/*japonica* cross in rice**. *Chinese J Rice Sci* 2000, **14**:71-78.

16. •• Zhao Q, Zhang Y, Cheng Z, Chen M, Wang S, Feng Q, Huang Y, Li Y, Tang Y, Zhou B *et al.*: **A fine physical map of the rice chromosome 4**. *Genome Res* 2002, **12**:817-823.
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.

17. The *Arabidopsis* Genome Initiative: **Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*.** *Nature* 2000, **408**:796-815.

18. Rossberg M, Theres K, Acarkan A, Herrero R, Schmitt T, Schumacher K, Schmitz G, Schmidt R: **Comparative sequence analysis reveals extensive microcolinearity in the lateral suppressor regions of the tomato, *Arabidopsis*, and *Capsella* genomes**. *Plant Cell* 2001, **13**:979-988.

19. Fu H, Dooner HK: **Intraspecific violation of genetic colinearity and its implications in maize**. *Proc Natl Acad Sci USA* 2002, **99**:9573-9578.

20. Song R, Llaca V, Messing J: **Mosaic organization of orthologous sequences in grass genomes**. *Genome Res* 2002, **12**:1549-1555.

21. Tarchini R, Biddle P, Wineland R, Tingey S, Rafalski A: **The complete sequence of 340 kb of DNA around the rice *Adh1-adh2* region reveals interrupted colinearity with maize chromosome 4**. *Plant Cell* 2000, **12**:381-391.

22. Goff SA: **Rice as a model for cereal genomics**. *Curr Opin Plant Biol* 1999, **2**:86-89.

23. Devos KM, Gale MD: **Genome relationships: the grass model in current research**. *Plant Cell* 2000, **12**:637-646.