International Rice Genome Sequencing Project: the effort to completely sequence the rice genome

Takuji Sasaki* and Benjamin Burr†

The International Rice Genome Sequencing Project (IRGSP) involves researchers from ten countries who are working to completely and accurately sequence the rice genome within a short period. Sequencing uses a map-based clone-by-clone shotgun strategy; shared bacterial artificial chromosome/P1-derived artificial chromosome libraries have been constructed from Oryza sativa ssp. japonica variety ‘Nipponbare’. End-sequencing, fingerprinting and marker-aided PCR screening are being used to make sequence-ready contigs. Annotated sequences are immediately released for public use and are made available with supplemental information at each IRGSP member’s website. The IRGSP works to promote the development of rice and cereal genomics in addition to producing genome sequence data.

Addresses
*Rice Genome Research Program, National Institute of Agrobiological Resources, 1–2, Kannondai 2-chome, Tsukuba, Ibaraki 305-8602, Japan; e-mail: tsasaki@abr.affrc.go.jp
†Biology Department, Brookhaven National Laboratory, Upton, New York 11973, USA; e-mail: burr@bnl.gov


0952-7915/00/$ – see front matter © 2000 Elsevier Science Ltd. All rights reserved.

Abbreviations
BAC bacterial artificial chromosome
EST expressed sequence tag
INE Integrated Rice Genome Explorer
IRGSP International Rice Genome Sequencing Project
PAC P1-derived artificial chromosome
YAC yeast artificial chromosome

Introduction
Rice is a wonderful plant. It feeds about one half of the world’s population, mainly in Asia, Africa, and South America. Cooking of rice is simple and does not require fermentation by yeast. It contains all of the amino acids essential for humans except lysine. It has a long cultivation history and, like religion or tradition, its use is deeply ingrained in the daily lives of Asian people.

A larger and more affluent population means, on the one hand, demand for greater production and better quality rice, and on the other hand, the availability of less land, water and labor to produce the crop. In short, there will be great demands on biotechnology to improve rice production.

The sequencing of all of the rice genes alone provides insufficient information on which to base crop improvements such as greater yield. Map-based sequence information is required to exploit the full potential of the rice sequence. In recent years, plant breeding has been enhanced by molecular-marker technology that permits researchers to screen larger populations and necessitates less progeny testing. Knowledge of the location of all of genes in a genome extends the usefulness of molecular-marker technology because it allows the identification of candidate genes that control specific traits. The genes themselves then become markers and the process becomes more accurate and efficient. For example, knowing the location and sequence of candidate genes makes it possible to design allele-specific markers; these markers readily lend themselves to processes in which the extraction of DNAs from plant leaves and the successive PCR reaction are automated.

Rice is a model species for the cereals and a good candidate for DNA sequencing. It has a genome size of 400–430 million base pairs (Mb), the smallest of the major cereals but three times that of Arabidopsis thaliana [1]. Rice also has a well-mapped genome: the rice molecular map, which has over 6000 markers, has already been useful in helping to align physical chromosome maps. Over 40,000 expressed sequence tags (ESTs) have been reported and many are mapped. A yeast artificial chromosome (YAC) library that has been fingerprinted and ordered with mapped markers currently covers 60% of the rice genome. Several bacterial artificial chromosome (BAC) libraries have also been described. Since the introduction of new methods for Agrobacterium tumefaciens transformation, rice has become the easiest of all cereal plants to transform genetically. This tool permits geneticists to complement mutations, or to confer dominant phenotypes to verify gene function.

Following-on from the past decade’s progress in understanding the molecular genetics of rice, an effort to sequence the whole rice genome has become a reality, beginning in Japan in 1997. Other countries with an interest in rice genomics decided to cooperate in this laborious but meaningful task, which became the International Rice Genome Sequencing Project (IRGSP). Here, we briefly review the strategy that has been adopted for the sequencing of the rice genome.
Mapping: the link between genomics and genetics

The genetic map has maintained its central importance as the basic tool that links information in the nucleotide sequence to phenotypic traits throughout the rice genome-sequencing project. The first step in understanding rice at the DNA level is to make a linkage map based on polymorphisms within DNA sequences, such as restriction fragment length polymorphisms (RFLPs), simple sequence repeats (SSRs) and cleaved amplified polymorphic sequences (CAPSs). More than ten rice genetic maps have been published so far; the one described by Harushima et al. [2]\(^*\) is the finest and most precise. The genetic markers positioned on these maps are indispensable for assembling the large DNA fragments that are selected from genomic libraries for sequencing, and for ascertaining the chromosomal locations of these fragments.

The next step in rice genome analysis is the construction of a genome-wide physical map. So far, only one physical map of the rice genome, which was assembled using YACs, has been published [3]. A revised map featuring more genetic markers is available at http://www.staff.or.jp/ Publicdata.html and is estimated to cover 60% of the genome. YACs have several disadvantages as templates for DNA sequencing, including chimerism and the difficulty of separating them from other yeast chromosomes. Therefore, BAC/PAC (P1-derived artificial chromosome) vectors are now increasingly used to construct new rice genomic libraries [4,5]. Several types of restriction enzymes, such as Sau3AI for PACs, and HindIII and EcoRI for BACs, have been used to address the uneven distribution of restriction sites in the rice genome. In particular, 37,000 clones in the BAC library using HindIII have been partially sequenced from both ends and have been fingerprinted to construct a BAC physical map [6]\(^*\). This method will result in a genome-wide physical map if the BACs in the library contain fragments from the whole of the rice genome: so far no information on this coverage is available.

Another strategy involves mapping many EST markers on the YAC physical map to generate a dense EST map [7]. More than 40,000 cDNA clones have been partially sequenced to generate ESTs [8]. These clones have subsequently been assembled based on their 3’-end sequences: by October 1999, about 5000 ESTs, each representing an independent group, had been mapped on the YAC physical map. These markers are thought to partly reflect the distribution of genes along each chromosome and should help in making sequence-ready contigs for gene-rich regions.

Sequence-ready PAC contigs have been identified by EST selection, and their order and degree of overlap has been confirmed by fingerprinting with by HindIII, EcoRV or BgII (T Baba, M Nakashima, T Sasaki, unpublished data). Sequencing several of the PACs chosen by this strategy indicates a higher gene density than expected: one gene every five thousand base pairs (5 kb) [9]\(^*\). It is estimated that the PAC contigs constructed in this way so far cover 30% of the genome. How can the gaps where no EST markers or YACs are available be filled? One strategy is to use information on the sequences that flank the gap to design PCR primers to ‘walk’ into the gaps. Another is to use end-sequence and fingerprint information from the BAC library to search for clones that fill the gap.

The completion of a deep, sequence-ready physical map, through the work outlined above is a prerequisite for the generation of a reliable map based on sequence data. This sequence-based map will be composed primarily of predicted genes that are supplemented with positional information from RFLP or EST markers, and phenotypic traits. It will become the skeleton of a database used for establishing the rice genome.

IRGSP sequencing strategy

Wide-ranging discussion within the IRGSP has encompassed many points including the optimal method of sequencing, the rice cultivar to be sequenced, the accuracy of sequences and the sequence release policy. A single variety of rice was chosen to be the source of DNA for sequencing because the cultivated varieties have diverse genetic backgrounds and so, if several varieties were used, allelic polymorphisms would probably impede the accurate compilation or integration of sequences. The japonica cultivar ‘Nipponbare’ (Figure 1) had already been used by the Rice Genome Research Program as a resource for extensive EST sequencing, and the construction of a dense linkage map and a YAC physical map. This work provides a valuable resource for genomic sequencing and we therefore decided to use Nipponbare as the common template throughout the IRGSP.

The members of the IRGSP have also chosen to accept common standards for sequence quality, annotation and sequence release. In the case of sequence quality and release policy, the IRGSP adopted the standards of the Human Genome Project [10], which sets a standard of less than one base-pair (bp) error in 10,000 bp. Although the level of accuracy is difficult to verify, this standard is achievable through a combination of high-quality shotgun sequence reads, a seven-fold redundancy and the insistence that 97% of all bases are sequenced on both strands or that two chemistries are used. The level of accuracy can be gauged by the use of established computational software, such as phred/phrap/consed (developed by P Green, B Ewing and D Gordon at the University of Washington). It can be experimentally verified by comparing the size of PCR products and restriction digests with predicted patterns based on sequence.

The IRGSP agreed to the immediate release of finished, but not necessarily annotated, sequence in units of intact BAC/PAC inserts to a public database such as the DNA Data Bank of Japan (DDBJ), GenBank or the European Molecular Biology Laboratory (EMBL). By January 2000,
22 Nipponbare PAC/BAC clones of a total of 3.3 Mb had been released. In addition, about 68,000 BAC-end sequences from Nipponbare libraries have been submitted to public databases.

**Annotation and database**

The sequences generated by the Rice Genome Research Program are annotated by searching the non-redundant protein database using BLASTX software [11], searching the rice EST database using BLASTN software [11], and scanning the sequence with GenScan [12] (trained for maize) to predict open reading frames and with Splice Predictor [13] to project exon/intron splice sites. These results are combined to make a final annotation of genes and elements, and their coordinates in a genome sequence. Other prediction tools such as Gene Finder [14], GeneMark [15], and NetPlantGene [16] are used by other groups.

Annotation is used not only for gene prediction but for characterization of sequences, such as repeats. Inverted and tandem repeats are predicted and drawn by Miropeats [17]. This tool is also very helpful in assembling the shotgun sequence. Transposable elements are recognized using gag and pol genes as references and then flanking long terminal repeats (LTRs) are identified.

A new genome database, named Integrated Rice Genome Explorer (INE), has been developed to integrate map and sequence information into a basic database for rice genomics [18••]. A web interface based on a Java applet allows rapid viewing of the database by scrolling and zooming. In this database, DNA markers on the genetic map play key roles in linking the YAC physical map, the EST map and the PAC/BAC physical map to define the chromosomal locations of clones on each map. Annotated genome sequences are available via PAC/BAC clones on a physical map, and the annotated information described above is shown in each table for predicted genes, ESTs assigned by BLASTN, proteins assigned by BLASTX, and so on. If a predicted gene shows significant similarity with a registered protein in a public database, detailed information on the registered gene or protein will soon be available by hyperlinks to the appropriate database. In the near future, physiological and biochemical data will be incorporated to link metabolic pathways to genomic information.

Recently, several important plant genes, linked to traits such as dwarfism and disease resistance, were discovered by map-based cloning and gene disruption [19–22]. Understanding these genes provides clues for elucidating the biochemical signal transduction pathways of plant hormones or pathogen-related defense mechanisms. Once the genes involved in any of these pathways are defined, INE will incorporate the results to add valuable information to the map and genome sequences.

Rice has important syntenic relationships with other cereal species [23•]. Similarities between the rice genome and those of other cereals will, therefore, be made apparent by linking INE to databases of each of the other important cereal crop species. A preliminary comparison is currently being tried between rice and maize databases [24].

---

**Table 1**

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Japan, Korea</td>
</tr>
<tr>
<td>2</td>
<td>United Kingdom (EU), Canada</td>
</tr>
<tr>
<td>3</td>
<td>USA</td>
</tr>
<tr>
<td>4</td>
<td>China ( indica variety Guang Lu Ai 4)</td>
</tr>
<tr>
<td>5</td>
<td>Taiwan</td>
</tr>
<tr>
<td>6</td>
<td>Japan</td>
</tr>
<tr>
<td>7</td>
<td>(Yet to be claimed)</td>
</tr>
<tr>
<td>8</td>
<td>(Yet to be claimed)</td>
</tr>
<tr>
<td>9</td>
<td>Thailand</td>
</tr>
<tr>
<td>10</td>
<td>USA</td>
</tr>
<tr>
<td>11</td>
<td>India, USA</td>
</tr>
<tr>
<td>12</td>
<td>France</td>
</tr>
</tbody>
</table>
Conclusions
The IRGSP grew out of a workshop held in 1997 at the 4th International Plant Molecular Biology Conference in Singapore. At this workshop, Japan, the USA, the European Union, Korea, and China agreed to collaborate on rice genome sequencing. Specifically, they agreed to share materials and results and to sequence Nipponbare as the sole germplasm. Since then, meetings have been held twice a year to report on progress and to discuss strategies and technical issues. The ten countries now taking part in the IRGSP are listed in Table 1. Detailed guidelines on joining IRGSP, meeting reports and an electronic journal, Oryza, are available at http://www.staff.or.jp/Seqcollab.html.

Rice sequencers realize that many scientists working in plant breeding, plant molecular genetics, plant molecular biology, and bioinformatics are awaiting the publication of the complete, high-quality rice genome sequence. The IRGSP aims to maintain high sequence quality standards, to publicize all sequence information as soon as possible and to complete the whole genome sequence in the shortest possible time. Intimate collaboration within IRGSP will make this plan realistic and will raise the profile of cereal genomics worldwide.

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


This map is still important in analyzing the rice genome because the probes, sequences and other information are publicly available on-line from the DNA bank at the Ministry of Agriculture, Forestry and Fisheries of Japan (URL http://bank.dna.affrc.go.jp/index.html).


6. Clemson University Genomics Institute: Rice BAC end sequencing project. URL http://www.genome.clemson.edu/projects/rice_bac_end.html


9. Rice Genome Research Program: Genome Sequencing. URL http://www.staff.or.jp/genomidata/GenomeFinished.html

• • On this web page, annotated and published rice PAC sequences are available via the INE. This tool is operable by Netscape Navigator with Java script (see also [18••]). Further information on the Rice Genome Research Program is available on the home page of this URL.


A new database tool commonly used in the IRGSP is introduced in this article. See also [9••].


This article describes synteny among grass species and stresses the importance of comparative genetics using rice as a model plant.