# 2 Rice Genome Sequence: The Foundation for Understanding the Genetic Systems

Takashi Matsumoto[1], Rod A. Wing[2], Bin Han[3], Takuji Sasaki[1]

*[1]National Institute of Agrobiological Sciences, 2-1-2 Kannondai, Tsukuba, Ibaraki 305-8602, Japan; [2]Department of Plant Sciences, The University of Arizona, Tucson, AZ 85721, USA; [3]National Center for Gene Research, Chinese Academy of Sciences, 500 Caobao Rd., 200233 Shanghai, China*

Reviewed by Satoshi Tabata



## 2.1 The Importance of the Accurate Genome Sequence of Rice

Progress in DNA sequencing technology has produced a tremendous increase in the number of nucleotide sequences from diverse organisms in a relatively short period of time. The collections of DNA and RNA sequences submitted to public databases such as GenBank, DDBJ, and EMBL recently reached 100 gigabases (NLM 2005) from 165,000 organisms. As sequencing advances, it is important to evaluate the accuracy and quality of the

sequence data. Positional accuracy indicates that the sequence is mapped onto the correct position on the genome. Sequence accuracy means that the nucleotide evaluation is performed correctly.

The first two sequencing technologies were the Maxam-Gilbert method (Gilbert and Maxam 1973) and Sanger dideoxy-chain terminator method (Sanger et al. 1977b). The Sanger method is widely used today because it is compatible with autosequencers that use fluorescent-labeled nucleotide analogs instead of radiolabeled chemicals (Smith et al. 1986).

The recent development of capillary sequencers that can simultaneously run 96 or 384 samples in 2 to 3 hours allows extensive parallel analysis of the nucleotide sequences. Improvements in liquid-handling robots and computer-aided data analysis technologies allow genome-wide sequencing in a reasonable time. The first "genome" sequence was that of a bacteriophage (Sanger et al. 1977a), followed by a bacterium (Fleischmann et al. 1995), and thereafter applied to the other organisms with larger genomes. Two major strategies have been devised for genome sequencing. In the hierarchical shotgun strategy, detailed, sequence-ready physical maps are constructed from genomic clones, and each clone such as P1-phage derived artificial chromosome (PAC), bacterial artificial chromosome (BAC), or cosmid, or fosmid clone is subcloned using partially digested DNA, and the subclones are sequenced (shotgun sequencing). Sequences are then assembled via sequence assembly software to form a contiguous sequence (contig) that virtually represents the original clone sequence. Finally, the clone sequences are connected according to the order of the physical maps to form the genome sequence. The strategy usually gives long, accurate sequences, although it is expensive in terms of time, monetary cost, and labor.

The alternative strategy, the whole genome shotgun (WGS) method (Venter et al. 1996), assembles the many short shotgun sequences derived from the whole genome to reconstruct the overall structure. The method is simple and straightforward, and is compatible with high-throughput sequencing equipment. The WGS method can supply genome-wide sequences mostly from "gene-rich" regions in a relatively short period of time. However, it usually gives many unconnected contigs. Moreover, there is a significant chance of genome misassembly in the case of repeat-rich sequences.

Choice of the genome sequencing strategy depends on the need. Obviously for the genome of a "model" organism that would become a key to the understanding of related species, one should aim for very high position and sequence accuracy so that it can serve as a reliable "reference" genome for subsequent comparisons with many other related organisms. On the other hand, analysis of an organism for a special purpose, such as to investigate genes involved in organism-specific metabolic pathways, requires only the genes involved in the pathways. Once the "reference"

genome is available, genomes of related organisms can be analyzed via the WGS method.

Rice is one of the most important crops worldwide, as it is the staple food for half the world's population. More than 2 billion people in Asia obtain the majority of their calories and protein from rice. As the world population continues to grow, as does the struggle to keep up the food supply, improving rice production is a pressing matter in the early 21st century. This makes rice the most economically and politically important crop.

Rice is also the key plant to understanding the genus *Oryza*, grass family (Gramineae) plants, and monocotyledons. *Oryza* is estimated to have originated 50 M years ago (Gaut 2002) and is represented by 23 species (Vaughan et al. 2003). Gramineae has approximately 10,000 species (Royal Botanic Gardens, Kew, http://www.rbgkew.org.uk/) and is the most ecologically and economically important of all the plant families. Colinearity of gene order (synteny) occurs across the grass family and many genes are mapped via this syntenic relationship. Rice is regarded as a "reference" crop that should be sequenced with as high an accuracy as possible. Accurate rice sequence information would be useful not only for isolation and breeding of the rice gene, but also for the molecular breeding of other important crops such as maize, barley, sorghum, and wheat. Researchers also recognize that revealing the rice genome drives the basic science of monocots, which cannot be fully understood from knowledge on *Arabidopsis* and other dicots (Leach et al. 2002).

## 2.2 Construction of the Sequence-Ready Physical Maps

In the hierarchical shotgun strategy, or "clone-by-clone methodology," large genomic DNA is digested into intermediate-sized fragments (40 to 150 kb), that are cloned into *E. coli* cells to make genome libraries. The libraries need to have enough redundancy in terms of both genome coverage and digestion sites. In the construction of IRGSP (the International Rice Genome Sequencing Project) Nipponbare physical maps, one PAC and three BAC libraries consisting of approximately 210,000 clones were first constructed (Baba et al. 2000; Mao et al. 2000). The libraries seem to cover all the rice genome because they have a 57.4× redundancy and have enough variety for restriction sites. Moreover, the Monsanto donated 3,416 BAC clones from their physical maps with the draft sequences to accelerate international attempts to complete the rice genome (Barry 2001). However, it was later revealed that the clones were still missing some part of the genome, leaving gaps in the physical maps.

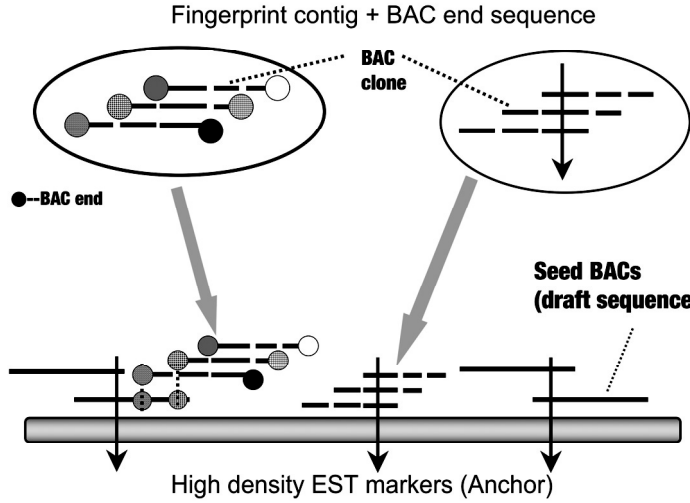


**Fig. 2.1.** A strategy for constructing the sequence-ready physical maps. The dotted lines indicate the fingerprinted BAC contigs, and the circles at the end of BACs show the BAC end sequences. Arrows crossing the BAC contigs indicate the anchor markers. The rectangle below shows the rice genome, to which the BAC contigs are mapped through anchor markers

The IRGSP took a complementary approach to make a comprehensive sequence-ready PAC/BAC physical map (Fig. 2.1). The Rice Genome Research Program (RGP) in Japan constructed a high-density transcript map in which 6,591 expressed sequence tag (EST)/sequence tagged site (STS) markers were mapped (Wu et al. 2002). An extensive, pooled clone polymerase chain reaction (PCR) screening identified the experimentally anchored PAC/BAC clones (Wu et al. 2003). Conversely, Clemson University Genomics Institute/Arizona Genomics Institute/Arizona Genomics Computational Laboratory (CUGI/AGI/AGCoL) from the USA fingerprinted and end-sequenced all BAC clones and assembled them into contigs via FingerPrinted Contigs (FPC) software (Soderlund et al. 1997). These assembled contigs were anchored to the genome by screening with the overlapping oligonucleotide (overgo) probes from the genetic markers

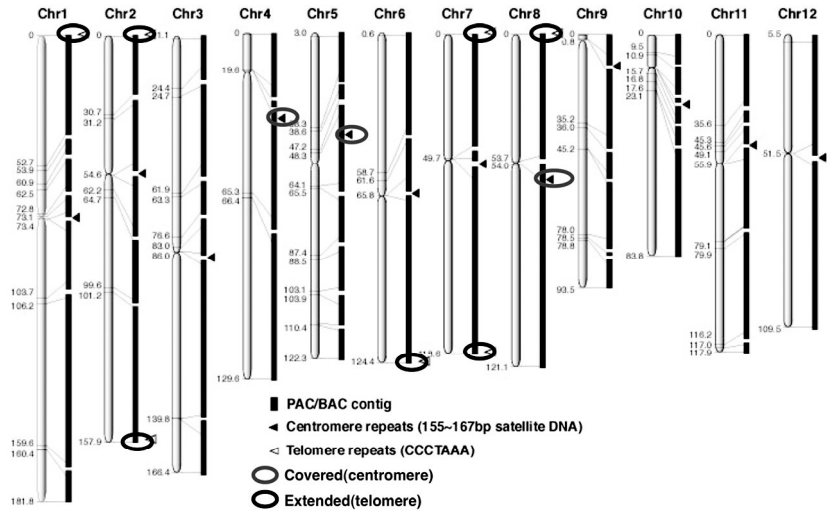


**Fig. 2.2.** Most recent status of Nipponbare physical maps. In each chromosome, linkage maps are shown on the left (numbers show the genetic distances) and PAC/BAC contigs (black bars) on the right. (Modified from the International Rice Genome Sequencing Project [2005] Nature 436:793–800)

(Chen et al. 2002). Finally, these two contrasting methodologies were combined to form a joint maps to finalize the physical mapping; from the draft sequences of "seed BACs" that are anchored by the DNA markers, the BAC end-sequence database was searched to detect the neighboring BAC clones, which are part of a contig. This sequence tagged connector (STC) method (Siegel et al. 1999) could effectively "walk" and "jump" between the marker-associated PAC/BAC contigs to fill the gaps. Eventually, most of the chromosomal clone gaps were successfully filled except for 36 remaining ones (Fig. 2.2, centromere region not counted). The sizes of all the remaining regions were measured by fiber-fluorescence *in situ* hybridization (FISH) analysis to be no longer than 100 kb. So far it is not known why these regions have not cloned. Several explanations (absence of available restriction digestion sites, sequence toxic to bacteria, complex repeat clusters that hamper clone identification by the DNA markers) are possible. There are also relatively large (0.2 to 2 Mb) gaps in the centromeric regions (black triangles in Fig. 2.2) for all but three chromosomes in the physical maps (chromosomes 4, 5, and 8). Even

considering these gaps, the IRGSP was able to construct a physical map covering more than 95% of the rice genome (see Table 2.1).

**Table 2.1.** Coverage of the IRGSP physical maps based on the sequenced length[a]

| Chromo-some | Sequenced length (Mb) | Gaps on arm regions | Centromere Covered | Estimated total (Mb) | Coverage (%) |
|---|---|---|---|---|---|
| 1 | 43.3 | 5 | No | 45.05 | 96 |
| 2 | 35.0 | 3 | No | 36.78 | 98 |
| 3 | 36.2 | 4 | No | 37.37 | 97 |
| 4 | 35.5 | 3 | Yes | 36.15 | 98 |
| 5 | 29.7 | 6 | Yes | 30.00 | 99 |
| 6 | 30.7 | 1 | No | 31.60 | 97 |
| 7 | 29.6 | 1 | No | 30.28 | 98 |
| 8 | 28.4 | 1 | Yes | 28.57 | 100 |
| 9 | 22.7 | 4 | No | 30.53 | 74 |
| 10 | 22.7 | 4 | No | 23.96 | 95 |
| 11 | 28.4 | 4 | No | 30.76 | 92 |
| 12 | 27.6 | 0 | No | 27.77 | 99 |
| All | 370.7 | 36 | | 388.82 | 95 |

[a]Modified from the International Rice Genome Sequencing Project (2005) Nature 436:793–800.

## 2.3 Two-Step Strategy for Completion of Rice Genome Sequencing

The IRGSP used the clone-by-clone method to obtain an accurate rice sequence and followed a two-step sequence publication in the public databases. The overall procedure for the genome sequencing is as follows. First, the target PAC/BAC DNA is purified and sheared into the two shotgun libraries (2 kb and 5 to 7 kb inserts). Both ends of approximately 1,000 subclones each are sequenced, and the subclone-end sequences are assembled via Phred-phrap software (http://www.phrap.org). Typically, one to five sequence contigs are formed from the resulting 4,000 shotgun sequences into a PAC/BAC clone (typically with a 100 to 150 kb genomic insert). As the sequence redundancy is high (>10×), most of the sequence gaps have multiple bridging shotgun clones, which make all the contigs both ordered and oriented. These sequence contigs can be submitted as phase 2 state in high-throughput genomic (HTG) sequences division of the public database of the National Centre for Biotechnology Information (NCBI, http://www.ncbi.nlm.nih.gov/projects/HTGS/).

The IRGSP decided to publish all the clone sequences from phase 2 or the high-quality draft of genome sequence because of the strong demand

for the release of relatively accurate genome sequences by crop researchers. The IRGSP constructed the pipelines for the mass-sequence production and submitted the clone-based sequences to the public databases immediately after the sequence assembly was completed. This accelerated data release resulted in the availability of a high-quality draft sequence of more than 450 Mb (366 Mb after removal of overlaps) by December 2002 (http://rgp.dna.affrc.go.jp/rgp/Dec18_NEWS.html).

The final step, converting the draft sequences into a continuous high-quality sequence, consists of three main parts: filling the sequence gaps, improving the low-accuracy regions, and resolving mis-assemblies. Filling the sequence gaps is a relatively easy task because all we need to do is fully sequence the bridge clones and reassemble the sequence. The IRGSP follows the Bermuda standard (http://www.genome.gov/ 10001812), which requires 99.99% accuracy for most of a finished sequence. The accuracy is evaluated by the scoring function of the Phred software. To improve the low-accuracy sequences indicated by low Phred scores, we resequenced them with different DNA polymerases or sequencing chemistries to reconfirm base determination. Although the rice genome has relatively few repeat sequences compared with other crops, every PAC/BAC clone sequence nonetheless has some transposon sequences, simple repeats, or unnamed repeats. Although these sequences are not genes for proteins, they might still have some unknown functions and be transcribed into RNAs or act as target sites for other proteins. As the assembly software combines the sequences by annealing similar regions, it has a high tendency to mis-assemble at these repeat regions. Trained researchers need to detect mis-assemblies, resolve them, edit the repeats to identify and order each repeats unit, and reassemble the sequences. Finally, the assembled sequences are verified by comparing sizes with those of real and virtual restriction digestion fragments. The finished sequences are submitted to the public databases as final HTG phase 3 or PLANT (PLN) sequences with or without annotations. At the time the draft sequencing was completed, more than 2000 PAC/BAC clones were left in phase 2, about half of which were assigned in RGP. The IRGSP continued working on finishing these sequences. Gradually the phase 2 clones became finished (PLN) clones, and all the sequencing was completed in December 2004 (Fig. 2.3). In the publication of the complete sequence, the IRGSP submitted 3,401 PAC/BAC clones, 18 fosmid clones, and some virtual clones from the sequences of PCR-amplified fragments. The total nucleotide length calculated by combining each PAC/BAC sequence and removing the overlaps is 370,733,456 bases. Adding these sequence lengths and the estimated gap lengths reveals the total physical length of

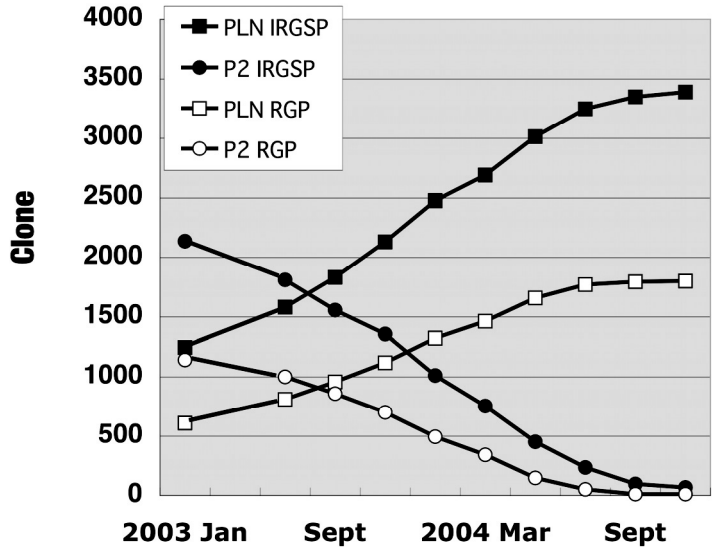


**Fig. 2.3.** Progress of finishing rice genome sequencing by IRGSP. P2 and PLN show phase 2 and completed clones, respectively

the rice genome to be 388.82 Mb. Three of the twelve centromeres have physical contigs, and two of them are published as high-quality sequences (chromosome 8: Wu et al. 2004; Nagaki et al. 2004; and chromosome 4: Zhang et al. 2004). These centromere sequences gave interesting materials for comparative genetics within the genome. Although the compositions of the two centromeres (CentO repeat, centromere retrotransposon of rice [CRR], and other transposons) are similar, the distributions of the CentO clusters are totally different. This suggests that chromosomes 4 and 8 have different histories of divergence (Ge et al. 1999). The sequenced regions, 370 Mb, correspond to 95.3% of the genome (98.9% in the euchromatin region). These results indicate that the IRGSP achieved the near-complete genome sequence of Nipponbare.

The high-quality and map-based sequence of the entire genome is now available in public databases. The Nipponbare genome sequence has been improved and published (http://rgp.dna.affrc.go.jp/IRGSP/ Build2, http://rgp.dna.affrc.go.jp/IRGSP/Build3, http://rgp.dna.affrc.go.jp/IRGSP/ Build4).

## 2.4 An Alternative Approach—the Whole Genome Shotgun Sequencing of Rice

Two activities have contributed to the whole genome shotgun rice sequencing. The Beijing Genomics Institute (BGI) has published the assembled 466-Mb sequence of *indica* variety, 93-11 from the 4× coverage WGS assembly (Yu et al. 2002). As described in a recent publication, this assembly was improved with the additional shotgun sequences (Yu et al. 2005). A private firm, Syngenta, also published 420 Mb of the Nipponbare sequence obtained by their independent WGS assembly (Goff et al. 2002). Both research groups have predicted 30,000 to 50,000 genes on the rice genome and also found many putative orthologs of genes from *Arabidopsis* or other plant species. Yu et al. (2005) have further improved the Syngenta Nipponbare WGS assembly by reassembling and combining the *japonica* and *indica* genome sequences.

### 2.4.1 Whole Genome Shotgun Sequencing of *japonica* Rice (Syngenta)

The latest assembly of Syngenta sequences by BGI assembled shotgun sequences (~6× coverage) of Nipponbare into 433.2-Mb sequences with 35,047 contigs (Yu et al. 2005). A total of 45,824 genes were predicted. Nearly 99% of the nonredundant rice full-length cDNA sequences (Kikuchi et al. 2003) showed corresponding sequences in the assembled genome.

### 2.4.2 Whole Genome Shotgun Sequencing of *indica* Rice (BGI)

The latest assembly of BGI assembled approximately 6.3× coverage shotgun sequences of *indica* cv. 93-11 into 466.3-Mb sequences with 50,233 contigs (Yu et al. 2005). A total of 49,088 genes were predicted and 97.1% of the nonredundant rice full-length cDNA sequences matched the assembled genome. Sequence comparison indicated 3.00 single nucleotide polymorphisms (SNPs) per kilobase in the genic regions and 15.13 SNPs/kb between Nipponbare and 93-11.

### 2.4.3 Comparison of Genome Sequences Derived from Whole Genome Shotgun Sequencing and Clone-by-Clone Shotgun Sequencing (IRGSP)

To compare the Syngenta and BGI shotgun sequences with the IRGSP map-based clone-by-clone sequence, we first mapped the BGI and

Syngenta contigs to IRGSP pseudomolecules using BLAST. With the Syngenta contigs we used the stringent condition that each contig must have at least 95% alignment with IRGSP pseudomolecules with an identity of at least 95%. Under this condition, a total of 26,007 contigs could be mapped to the pseudomolecules covering 290 Mb, with coverage varying from 77% to 81%. Discrepancy of this result from cDNA mapping might be due to the sequence mis-assemblies in the repeat-rich region.

With the BGI contigs, considering the sequence variation between the two subspecies, we used the condition that each contig must align at least 50% with the IRGSP pseudomolecules and have at least 80% identity. Under this condition, we mapped a total of 25,101 contigs to pseudomolecules covering 258 Mb, with coverage varying from 58% to 78%, indicating subspecies variation derived from large insertions, deletions, and inversions.

As the sequence assembly obtained by the shotgun sequencing is inherently confusable with repetitive sequences, we also analyzed the shotgun sequence coverage in genic regions. We used the dataset of 37,544 of IRGSP predicted genes, among which 9,485 genes are supported by rice transcripts. Of these predicted genes, 26,424 (70%) were fully covered by the Syngenta contigs and 22,376 (60%) were fully covered by BGI contigs. In full-length cDNA supported genes, Syngenta contigs covered 7,139 (75%) and BGI contigs covered 6,482 (68.3%), which may reflect the relative high coverage in the gene-dense regions compared to other parts of the genome.

Detailed study of a region of chromosome 1 shows that each assembly contains nonhomologous, misaligned, or duplicate coverage, which may be an artifact of the assembly program. It even shows 0.05% base-pair mismatches in matched contigs within Nipponbare, possibly a result of the relatively low coverage of shotgun sequences. A case study of the CentO repeat sequences showed that 32% of this centromere-specific sequence was found in contigs outside the centromere, indicating a high rate of mis-assembly of the WGS with repeat sequences.

## 2.5 Initial Analysis of the Rice Genome

After completion of sequencing, the IRGSP presented the results of the initial analysis (IRGSP 2005). About one-third of the total genome contains the known repeat elements, including transposons. In repeat-free domains, the computer program FGENESH detected 37,544 protein-coding sequences, 60% of which have some similarities to rice ESTs and cDNAs. Seventy percent of the predicted genes have at least one homolog in the *Arabidopsis* proteome. About 2,800 gene models that match rice transcripts have no counterpart in *Arabidopsis* detected by BLASTP with a cutoff value of $10^{-20}$, and most of these proteins have no known function.

About 30% of the predicted genes are present in tandem duplications. A graphical presentation of distribution of the major gene clusters on each chromosome is shown in Fig. 2.4. Apparently, there are many tandemly arrayed gene clusters (the pixels indicate each gene; the stacked pixels indicate that tandem array) in more than half of the 12 chromosomes. For example, the major cluster in chromosome 1 (extreme left) is a protein kinase cluster, and the gene cluster in chromosome 11 (extreme right) is related to disease resistance.

As rice is the crop plant that is widely utilized as the staple, much analysis was devoted to identifying some useful DNA markers, including more than 10,000 *Tos17* insertion sites, 19,000 class I simple sequence repeats (SSRs) sites, and 80,000 SNPs. These will be good candidates for the polymorphic markers among varieties and subspecies that would assist map-based cloning and marker-assisted selection.

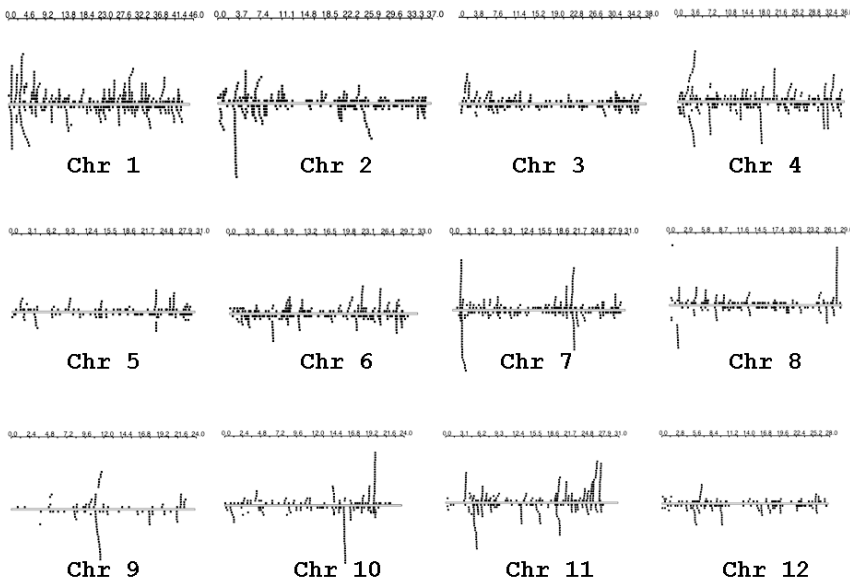


**Fig. 2.4.** Distribution of arrayed genes on rice genome. Only tandemly repeated genes were considered. A BLASTP search was performed within each chromosome against all predicted protein sequences by IRGSP. Proteins that have an expectation (*E*)-value of $10^{-5}$ among others were grouped and shown as pixels in the figure. Graphics were made through GenomePixelizer (http://niblrrs.ucdavis.edu/GenomePixelizer/GenomePixelizer_Welcome.html). Numbers above each plot show the positions in pseudomolecules (in Mb)

## 2.6 Current Status and Future Developments

After completion of the official tasks of the IRGSP, the member countries are trying to fill the remaining gaps and to improve their sequences. Telomere regions, which are considered to be responsible for accurate chromosome replication and maintenance, have not been represented in the PAC/BAC libraries. The telomere regions have specific structures of TTTAGGG repeats (Richards and Ausubel 1988) and few restriction digestion sites. A new genomic library with a fosmid vector has been produced by Arizona Genomics Institute (Ammiraju et al. 2005). This library, which has 110,592 clones with an average insert size of 41 kb, has been constructed via random physical shearing of the genfome, and it has been most helpful for the IRGSP in finding new clones to fill the clone gaps. Moreover, seven fosmid clones were recently found by hybridization with the unique probe sequences at the end of the chromosomes (positions of these telomere clones are indicated by encircled white triangles in Fig. 2.2). All these clones have TTTAGGG repeats or its derivatives, indicating the physical contigs reside very near the ends of the chromosomes. The transition regions between euchromatin and telomere regions have approximately 600 predicted genes for seven subtelomeric regions. Searches for clones in other telomere regions are underway (Mizuno et al. 2006). These and other improvements are included in the updated version of rice pseudomolecules (currently build 4) at the IRGSP Web site.

The rice genome sequence is a milestone in understanding the grass family. Comparative mapping could be a useful tool for isolating the genes among other grasses. High-resolution comparative maps have been constructed between rice and wheat (Sorrells et al. 2003) and rice and maize (Salse et al. 2004). Several trait genes (*VRN1*: Yan et al. 2003; *VRN2*: Yan et al. 2004; *Ppd-H1*: Turner et al. 2005; *Ph1*: Griffiths et al. 2006) have been isolated from barley and wheat, and both syntenic as large genome blocks and microsyntenic relationships with the rice genome have been effectively utilized in gene mapping. For the “Crop Circle” investigators (Devos 2005), rice is regarded as a stepping-stone to finding the order of markers and genes in the larger genomes. Such syntenic mapping has been used in *Brassica* genomes, for which *Arabidopsis* is the standard, and in *Lotus japonicus* and *Medicago truncatula*, which serve as models for legume crops.

The rice genome sequence is the key to understanding the science of rice (Paterson et al. 2005). Knowledge of what constitutes rice, how rice grows and develops, how it produces grains, and how it resists pests and diseases can lead to more design-based agriculture and biotechnology. The new technologies will be the foundation to a second “Green Revolution” to allow sustainable growth of human life.

## Acknowledgments

The authors thank all the participants of the IRGSP. We also acknowledge all the rice biologists who have joined the analysis of feature of rice genome. We also thank to Dr. S. Tabata from Kazusa DNA Research Institute for a critical review of the manuscript.

## References

Ammiraju JS, Yu Y, Luo M, Kudrna D, Kim H, Goicoechea JL, Katayose Y, Matsumoto T, Wu J, Sasaki T, Wing RA (2005) Random sheared fosmid library as a new genomic tool to accelerate complete finishing of rice (*Oryza sativa* spp. Nipponbare) genome sequence: sequencing of gap-specific fosmid clones uncovers new euchromatic portions of the genome. Theor Appl Genet 111:1596–1607

Baba T, Katagiri S, Tanoue H, Tanaka R, Chiden Y, Saji S, Hamada M, Nakashima M, Okamoto M, Hayashi M, Yoshiki S, Karasawa W, Honda M, Ichikawa Y, Arita K, Ikeno M, Ohta T, Umehara Y, Matsumoto T, de Jong PJ, Sasaki T (2000) Construction and characterization of rice genomic libraries: PAC library of *japonica* variety, Nipponbare, and BAC library of *indica* variety, Kasalath. Bull Natl Inst Agrobiol Res Jpn 14:41–51

Barry GF (2001) The use of the Monsanto draft rice genome sequence in research. Plant Physiol 125:1164–1165

Chen M, Presting G, Barbazuk WB, Goicoechea JL, Blackmon B, Fang G, Kim H, Frisch D, Yu Y, Sun S, Higingbottom S, Phimphilai J, Phimphilai D, Thurmond S, Gaudette B, Li P, Liu J, Hatfield J, Main D, Farrar K, Henderson C, Barnett L, Costa R, Williams B, Walser S, Atkins M, Hall C, Budiman MA, Tomkins JP, Luo M, Bancroft I, Salse J, Regad F, Mohapatra T, Singh NK, Tyagi AK, Soderlund C, Dean RA, Wing RA (2002) An integrated physical and genetic map of the rice genome. Plant Cell 14:537–545

Devos KM (2005) Updating the “crop circle”. Curr Opin Plant Biol 8:155–62

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM, McKenney K, Sutton G, Fitzhugh W, Fields C, Gocayne JD, Scott J, Shirley R, Liu L, Glodek A, Kelley JM, Weidman JF, Phillips CA, Spriggs T, Hedblom E, Cotton MD, Utterback TR, Hanna MC, Nguyen DT, Saudek DM, Brandon RC, Fine LD, Fritchman JL, Fuhrmann JL, Geoghagen NSM, Gnehm CL, McDonald LA, Small KV, Fraser CM, Smith HO, Venter JC (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269: 496–512

Gaut B (2002) Evolutionally dynamics of grass genomes. New Phytologist 154:15–28

Ge S, Sang T, Lu BR, Hong DY (1999) Phylogeny of rice genomes with emphasis on origins of allotetraploid species. Proc Natl Acad Sci USA 96:14400–14405

Gilbert W, Maxam A (1973) The nucleotide sequence of the *lac* operator. Proc Natl Acad Sci USA 70:3581–3584

Goff SA, Ricke D, Lan TH, Presting G, Wang R, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H, Hadley D, Hutchison D, Martin C, Katagiri F, Lange BM, Moughamer T, Xia Y, Budworth P, Zhong J, Miguel T, Paszkowski U, Zhang S, Colbert M, Sun WL, Chen L, Cooper B, Park S, Wood TC, Mao L, Quail P, Wing R, Dean R, Yu Y, Zharkikh A, Shen R, Sahasrabudhe S, Thomas A, Cannings R, Gutin A, Pruss D, Reid J, Tavtigian S, Mitchell J, Eldredge G, Scholl T, Miller RM, Bhatnagar S, Adey N, Rubano T, Tusneem N, Robinson R, Feldhaus J, Macalma T, Oliphant A, Briggs S (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica)*. Science 296:92–100

Griffiths S, Sharp R, Foote TN, Bertin I, Wanous M, Reader S, Colas I, Moore G (2006) Molecular characterization of *Ph1* as a major chromosome pairing locus in polyploid wheat. Nature 439:749–752

International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. Nature 436:793–800

Kikuchi S, Satoh K, Nagata T, Kawagashira N, Doi K, Kishimoto N, Yazaki J, Ishikawa M, Yamada H, Ooka H, Hotta I, Kojima K, Namiki T, Ohneda E, Yahagi W, Suzuki K, Li CJ, Ohtsuki K, Shishiki T, Otomo Y, Murakami K, Iida Y, Sugano S, Fujimura T, Suzuki Y, Tsunoda Y, Kurosaki T, Kodama T, Masuda H, Kobayashi M, Xie Q, Lu M, Narikawa R, Sugiyama A, Mizuno K, Yokomizo S, Niikura J, Ikeda R, Ishibiki J, Kawamata M, Yoshimura A, Miura J, Kusumegi T, Oka M, Ryu R, Ueda M, Matsubara K, Kawai J, Carninci P, Adachi J, Aizawa K, Arakawa T, Fukuda S, Hara A, Hashizume W, Hayatsu N, Imotani K, Ishii Y, Itoh M, Kagawa I, Kondo S, Konno H, Miyazaki A, Osato N, Ota Y, Saito R, Sasaki D, Sato K, Shibata K, Shinagawa A, Shiraki T, Yoshino M, Hayashizaki Y, Yasunishi A (2003) Collection, mapping, and annotation of over 28,000 cDNA clones from *japonica* rice. Science 301:376–379

Leach J, McCouch S, Slezak T, Sasaki T, Wessler S (2002) Why finishing the rice genome matters. Science 296:45

Mao L, Wood TC, Yu Y, Budiman MA, Tomkins J, Woo S, Sasinowski M, Presting G, Frisch D, Goff S, Dean RA, Wing RA (2000) Rice transposable elements: a survey of 73,000 sequence-tagged-connectors. Genome Res 10:982–990

Mizuno H, Wu J, Kanamori H, Fujisawa M, Namiki N, Saji S, Katagiri S, Katayose Y, Sasaki T, Matsumoto T (2006) Sequencing and characterization of telomere and subtelomere regions on rice chromosomes 1S, 2S, 2L, 6L, 7S, 7L and 8S. Plant J 46:206–217

Nagaki K, Cheng Z, Ouyang S, Talbert PB, Kim M, Jones KM, Henikoff S, Buell CR, Jiang J (2004) Sequencing of a rice centromere uncovers active genes. Nat Genet 36:138–145

National Library of Medicine (2005) Public Collections of DNA and RNA Sequence Reach 100 Gigabases. Press Release, http://www.nlm.nih.gov/news/press_releases/dna_rna_100_gig.html

Paterson AH, Freeling M, Sasaki T (2005) Grains of knowledge: genomics of model cereals. Genome Res 15:1643–1650

Richards EJ, Ausubel FM (1988) Isolation of a higher eukaryotic telomere from *Arabidopsis thaliana*. Cell 53:127–136

Salse J, Piegu B, Cooke R, Delseny M (2004) New *in silico* insight into the synteny between rice (*Oryza sativa* L.) and maize (*Zea mays* L.) highlights reshuffling and identifies new duplications in the rice genome. Plant J 38:396–409

Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes CA, Hutchison CA, Slocombe PM, Smith M (1977a) Nucleotide sequence of bacteriophage phi X174 DNA. Nature 265:687–695

Sanger F, Nicklen S, Coulson AR (1977b) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 74:5463–5467

Siegel AF, Trask B, Roach JC, Mahairas GG, Hood L, van den Engh G (1999) Analysis of sequence-tagged-connector strategies for DNA sequencing. Genome Res 9:297–307

Smith LM, Sanders JZ, Kaiser RJ, Hughes P, Dodd C, Connell CR, Heiner C, Kent SB, Hood LE (1986) Fluorescence detection in automated DNA sequence analysis. Nature 321:674–679

Soderlund C, Longden I, Mott R (1997) FPC: a system for building contigs from restriction fingerprinted clones. Comput Appl Biosci 13:523–535

Sorrells ME, La Rota M, Bermudez-Kandianis CE, Greene RA, Kantety R, Munkvold JD, Miftahudin, Mahmoud A, Ma X, Gustafson PJ, Qi LL, Echalier B, Gill BS, Matthews DE, Lazo GR, Chao S, Anderson OD, Edwards H, Linkiewicz AM, Dubcovsky J, Akhunov ED, Dvorak J, Zhang D, Nguyen HT, Peng J, Lapitan NL, Gonzalez-Hernandez JL, Anderson JA, Hossain K, Kalavacharla V, Kianian SF, Choi DW, Close TJ, Dilbirligi M, Gill KS, Steber C, Walker-Simmons MK, McGuire PE, Qualset CO (2003) Comparative DNA sequence analysis of wheat and rice genomes. Genome Res 13:1818–1827

Turner A, Beales J, Faure S, Dunford RP, Laurie DA (2005) The pseudo-response regulator *Ppd-H1* provides adaptation to photoperiod in barley. Science 310:1031–1034

Vaughan DA, Morishima H, Kadowaki K (2003) Diversity in the *Oryza* genus. Curr Opin Plant Biol 6:139–146

Venter JC, Smith HO, Hood L (1996) A new strategy for genome sequencing. Nature 381:364–366

Wu J, Maehara T, Shimokawa T, Yamamoto S, Harada C, Takazaki Y, Ono N, Mukai Y, Koike K, Yazaki J, Fujii F, Shomura A, Ando T, Kono I, Waki K, Yamamoto K, Yano M, Matsumoto T, Sasaki T (2002) A comprehensive rice transcript map containing 6591 expressed sequence tag sites. Plant Cell 14:525–535

Wu J, Mizuno H, Hayashi-Tsugane M, Ito Y, Chiden Y, Fujisawa M, Katagiri S, Saji S, Yoshiki S, Karasawa W, Yoshihara R, Hayashi A, Kobayashi H, Ito K, Hamada M, Okamoto M, Ikeno M, Ichikawa Y, Katayose Y, Yano M, Matsumoto T, Sasaki T (2003) Physical maps and recombination frequency of six rice chromosomes. Plant J 36:720–730

Wu J, Yamagata H, Hayashi-Tsugane M, Hijishita S, Fujisawa M, Shibata M, Ito Y, Nakamura M, Sakaguchi M, Yoshihara R, Kobayashi H, Ito K, Karasawa W, Yamamoto M, Saji S, Katagiri S, Kanamori H, Namiki N, Katayose Y, Matsumoto T, Sasaki T (2004) Composition and structure of the centromeric region of rice chromosome 8. Plant Cell 16:967–976

Yan L, Loukoianov A, Tranquilli G, Helguera M, Fahima T, Dubcovsky J (2003) Positional cloning of the wheat vernalization gene *VRN1*. Proc Natl Acad Sci USA 100:6263–6268

Yan L, Loukoianov A, Blechl A, Tranquilli G, Ramakrishna W, SanMiguel P, Bennetzen JL, Echenique V, Dubcovsky J (2004) The wheat *VRN2* gene is a flowering repressor down-regulated by vernalization. Science 303:1640–1644

Yu J, Hu S, Wang J, Wong GK, Li S, Liu B, Deng Y, Dai L, Zhou Y, Zhang X, Cao M, Liu J, Sun J, Tang J, Chen Y, Huang X, Lin W, Ye C, Tong W, Cong L, Geng J, Han Y, Li L, Li W, Hu G, Huang X, Li W, Li J, Liu Z, Li L, Liu J, Qi Q, Liu J, Li L, Li T, Wang X, Lu H, Wu T, Zhu M, Ni P, Han H, Dong W, Ren X, Feng X, Cui P, Li X, Wang H, Xu X, Zhai W, Xu Z, Zhang J, He S, Zhang J, Xu J, Zhang K, Zheng X, Dong J, Zeng W, Tao L, Ye J, Tan J, Ren X, Chen X, He J, Liu D, Tian W, Tian C, Xia H, Bao Q, Li G, Gao H, Cao T, Wang J, Zhao W, Li P, Chen W, Wang X, Zhang Y, Hu J, Wang J, Liu S, Yang J, Zhang G, Xiong Y, Li Z, Mao L, Zhou C, Zhu Z, Chen R, Hao B, Zheng W, Chen S, Guo W, Li G, Liu S, Tao M, Wang J, Zhu L, Yuan L, Yang H (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). Science 296:79–92

Yu J, Wang J, Lin W, Li S, Li H, Zhou J, Ni P, Dong W, Hu S, Zeng C, Zhang J, Zhang Y, Li R, Xu Z, Li S, Li X, Zheng H, Cong L, Lin L, Yin J, Geng J, Li G, Shi J, Liu J, Lv H, Li J, Wang J, Deng Y, Ran L, Shi X, Wang X, Wu Q, Li C, Ren X, Wang J, Wang X, Li D, Liu D, Zhang X, Ji Z, Zhao W, Sun Y, Zhang Z, Bao J, Han Y, Dong L, Ji J, Chen P, Wu S, Liu J, Xiao Y, Bu D, Tan J, Yang L, Ye C, Zhang J, Xu J, Zhou Y, Yu Y, Zhang B, Zhuang S, Wei H, Liu B, Lei M, Yu H, Li Y, Xu H, Wei S, He X, Fang L, Zhang Z, Zhang Y, Huang X, Su Z, Tong W, Li J, Tong Z, Li S, Ye J, Wang L, Fang L, Lei T, Chen C, Chen H, Xu Z, Li H, Huang H, Zhang F, Xu H, Li N, Zhao C, Li S, Dong L, Huang Y, Li L, Xi Y, Qi Q, Li W, Zhang B, Hu W, Zhang Y, Tian X, Jiao Y, Liang X, Jin J, Gao L, Zheng W, Hao B, Liu S, Wang W, Yuan L, Cao M, McDermott J, Samudrala R, Wang J, Wong GK, Yang H (2005) The Genomes of *Oryza sativa*: a history of duplications. PLoS Biol 3:e38

Zhang Y, Huang Y, Zhang L, Li Y, Lu T, Lu Y, Feng Q, Zhao Q, Cheng Z, Xue Y, Wing RA, Han B (2004) Structural features of the rice chromosome 4 centromere. Nucl Acids Res 32:2023–2030

http://www.springer.com/978-0-387-48903-2

Rice Functional Genomics
Challenges, Progress and Prospects
Upadhyaya, N.M. (Ed.)
2007, XXXIV, 500 p. 59 illus., Hardcover
ISBN: 978-0-387-48903-2