JARQ 40 (2), 99 – 105 (2006) http://www.jircas.affrc.go.jp

REVIEW

Completion of Rice Genome Sequencing
—a Paradigm Shift of Rice Biology—

Takashi MATSUMOTO[1]*, Jianzhong WU[1], Nobukazu NAMIKI[2], Hiroyuki KANAMORI[2], Masaki FUJISAWA[2] and Takuji SASAKI[1]

[1] Department of Genome Research, National Institute of Agrobiological Sciences (Tsukuba, Ibaraki 305–8602, Japan)
[2] Research Division I, STAFF Institute (Tsukuba, Ibaraki 305–0854, Japan)

Abstract

The ultimate goal of the international rice genome sequencing consortium has been successfully realized with the completion of the entire rice genome sequence in 2004. An accurate sequence totaling 370 Mbp of the *japonica* rice cultivar Nipponbare corresponding to 95% of the entire genome is now publicly available. Comprehensive analysis of the sequence has revealed some unique features of the rice genome. Together with other rice genomics resources, this map-based sequence will revolutionize rice biology with the rapid identification of genes responsible for many agronomic traits, genome-wide comparison with other rice species and cereal crops, and establishment of new breeding methods based on DNA markers.

Discipline: Genetic resources
Additional key words: comparative genomics, gene annotation, high-throughput sequencing

Introduction

Rice is one of the most important crops in the world and is consumed by more than half the world population. About 20% of the total calorie supply worldwide comes from rice, and especially in Asia, more than 2 billion people derive 60–70% of their daily energy requirement from rice and its derivatives. The world population is increasing at 1–2% per year, but crop production does not increase proportionally. Recognition of the importance of rice as the world's foremost staple crop has prompted the United Nations to designate the year 2004 as the International Year of Rice under the slogan of "Rice is Life" (http://www.fao.org/rice2004/). Rice research, however, will not only benefit rice. The conservation of gene order among grass genomes[6,21] clearly indicates that the results of rice genome analysis could be well utilized for analysis of other grass crops such as millet, sorghum, sugarcane, maize, and wheat.

The Rice Genome Research Program (RGP) has established the genomic resources such as EST sequences and YAC framework physical maps as well as the high-density genetic maps and markers from 1991[26]. These maps and sequences served as the framework for establishing the map-based sequence of rice. All biological processes involved in the life cycle of rice such as fertilization, germination, growth, development, photosynthesis, metabolism, and response to the environment are encoded in the genome. Obtaining the master plan (genome sequence) of the rice plant is therefore indispensable in understanding rice biology. For these reasons, rice has long been a target plant for a comprehensive genome analysis. However, although many researchers recognized the impact of this undertaking, it is such an enormous task for one research institution or even a single country. The International Rice Genome Sequencing Project (IRGSP), composed of 10 countries, was therefore established in 1997 to pool resources and manpower, accelerate the sequencing of the genome and insure public access to the sequence data. After seven years of global collaboration, it has successfully attained its goal and has released an accurate Nipponbare genome sequence of 370 Mbp. In this paper, the sequencing effort of the RGP from draft sequencing to the completion of the sequence, as well as the impact of

*Corresponding author: fax +81–29–838–7468; e-mail mat@nias.affrc.go.jp
Received 14 October 2005; accepted 7 November 2005.

the accurate rice genome sequence to crop science and agriculture both at present and in the future will be described.

Rice genome sequencing by the IRGSP

1. Physical map construction

The IRGSP sets the accurate rice genome sequence as the ultimate goal (http://rgp.dna.affrc.go.jp/IRGSP/bnl/Guidelines.html) of this project. The word "accurate" has two meanings: (1) to identify each specific nucleotide with high quality; and (2) to ensure that the sequence is positionally correct. To achieve this goal, the clone-by-clone methodology coupled with the hierarchical shotgun sequencing was chosen. This strategy requires a sequence-ready physical map that covers the whole genome with genomic clones. The IRGSP constructed several genomic libraries with PAC[1] and BAC (RGP, unpublished or Mao et al.[19]) clones with more than 68 times genome coverage.

Two methods which complemented each other were used to construct physical contigs with a relatively high genome-coverage. The RGP constructed a transcript map with 6,591 EST markers which were then used to screen the PAC and BAC clones. Fingerprint contigs were constructed by the US group using the restriction digest fingerprints and the FPC software[29] that facilitated alignment of clone-contigs based on the similarity in the fragment patterns. With these combined strategies, the RGP succeeded in the construction of the physical maps of the 6 chromosomes (chr 1, 2, 6, 7, 8, and 9) in early 2003[37]. The remaining clone gaps were filled by "walking" from the clones at the contig ends, and the "Sequence-Tagged-Connector" method[18] in which blasting the gap-proximal sequence against the BAC end-sequence database automatically identified the contig-extending clones. The unsequenced regions consist of 36 gaps in the euchromatin, 9 gaps in the centromere regions, and the chromosome ends (telomere region) even if some BAC or fosmid clones extended into the specific repeat sequences that correspond to the telomeres. To determine the total physical size of the rice genome, the contig-gap length was measured by the fiber-FISH (fluorescence *in situ* hybridization) method[4]. Although most of the physical gaps were found to be relatively small (10–90 kb in size), PAC or BAC clones were not obtained even by the extensive screening of the whole libraries. This may imply that some regions in the genome may be unstable in *E. coli* and could not be easily cloned. In total, 3,453 clones of PACs, BACs, fosmids, and other artificial sequences were aligned to the 12 rice chromosomes.

2. Clone-based sequencing strategy

Two major methodologies are now widely used for genome sequencing: the hierarchical shotgun sequencing strategy and the whole-genome shotgun sequencing strategy[11]. In the former, an individual PAC/BAC clone is shotgun-sequenced and locally assembled[8] so that mis-assembly could be significantly minimized. In the latter, shotgun clones from the whole genome are sequenced and globally assembled by the high-throughput assembly software to reconstruct the genome sequences. The IRGSP adopted the hierarchical shotgun sequencing strategy although it requires several years of extra work to map the clone especially for a large genome such as rice. Additionally, the RGP constructed an efficient draft-sequencing pipeline which facilitated the completion of the high-quality draft sequence of the entire genome in December 2002 (http://rgp.dna.affrc.go.jp/rgp/Dec18_NEWS.html).

Completion of sequencing from the draft status involved several "finishing" protocols such as (1) gap-filling, (2) improving sequence quality to the IRGSP publication standard (99.99% accuracy), and (3) resolving mis-assembly. The sequence gaps were filled by full-sequencing of bridge clones and the low quality regions were re-sequenced using other chemistries and polymerases[24]. Mis-assembly was resolved by manual detection of the degenerated part of the sequence, selection of a representative subclone, and full sequencing of the subclone[28]. The sequencing pipeline was further modified with emphasis on manual editing and high-throughput re-sequencing of the regions to improve the sequence quality. These modifications increased the rate of finishing to 80–90 PAC/BACs per month, which corresponds to approximately 8 Mbp of genome sequence per month. During the same period, the RGP published the completed whole chromosomal sequence of chromosome 1[27] and the group from China (NCGR) published the sequence of chromosome 4[10]. To date, two other publications from the US group have been published[32,33].

The completion of rice genome sequencing was officially declared in December 2004. The total assembled nucleotide sequence is 370.7 Mbp (Build 3.0 Pseudomolecules). Adding the estimated size of gaps, the entire rice genome is estimated to be 388.82 Mbp in total. This is at present the most accurate size estimation of rice (*japonica*) mostly based on calculations from the nucleotide sequences. This estimation also indicates that the IRGSP revealed 95% of the total genome or 98.9% of the euchromatin region. The remaining 5% exists mostly as centromere regions and heterochromatin regions (Fig. 1). This indicates that we have obtained most of the genic information of the rice plant. The assembled genome

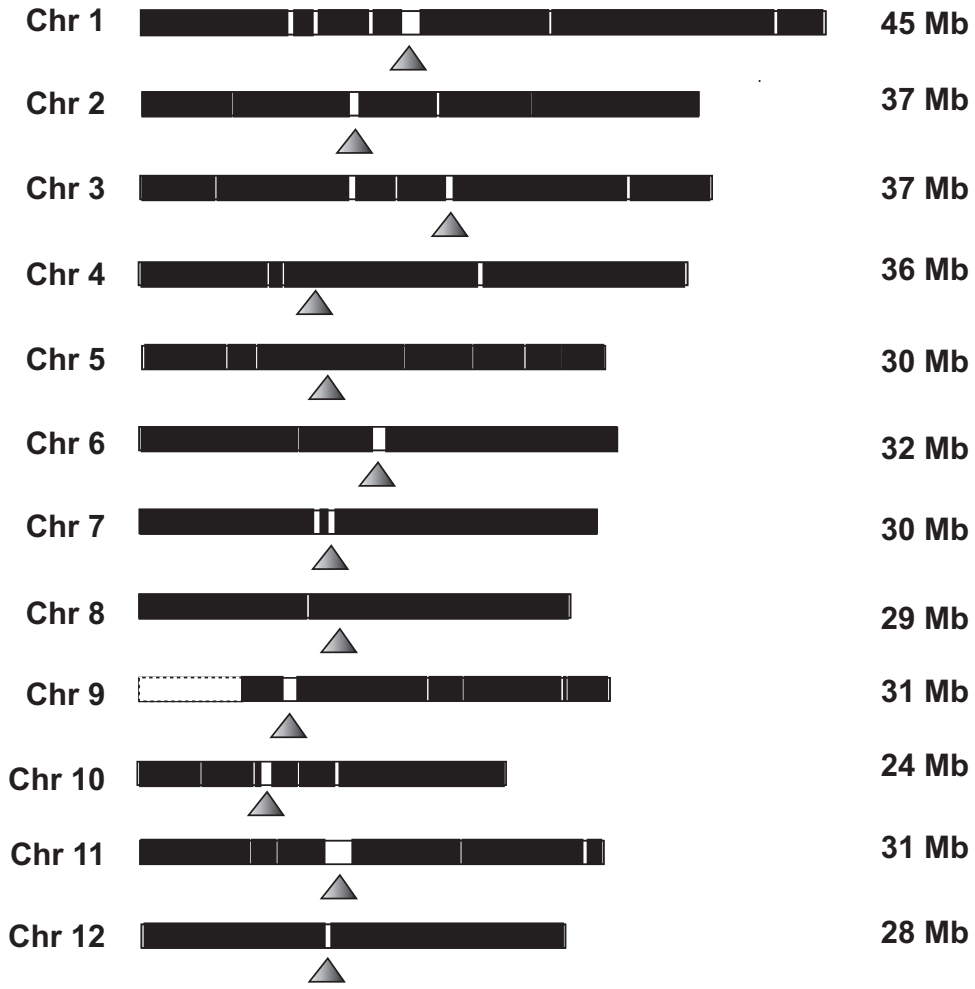

Fig. 1. Sequenced regions of the 12 rice chromosomes

For each chromosome, the length of the bar corresponds to the total physical size in kb which is also numerically shown in the right. Black boxes show the sequenced regions, while blank areas correspond to the unsequenced regions because of lack of clone contigs. The triangles in the middle indicate the approximate positions of the centromeres.

sequence used in the analysis described above (http://rgp.dna.affrc.go.jp/IRGSP/Build3/build3.html) and its updated version (Build 4.0 Pseudomolecules, http://rgp.dna.affrc.go.jp/IRGSP/Build4/build4.html) could be accessed from the IRGSP website (Fig. 2). From then on, the IRGSP has continued to analyze the unsequenced heterochromatin regions. The sequences of the centromere regions of chromosomes 8 and 4 have been revealed and analyzed[22,38,40]. Sequencing of the ribosomal gene cluster region and telomere regions is also in progress.

3. Analysis of rice genome

A comprehensive analysis of the completed genome sequence has been published recently[13]. Some remarkable characteristics of the rice genome are the following:

(1) Predicted genes. A total of 37,544 genes have been predicted in the rice genome. This is the largest number of genes identified in plant and animal genomes that have been completely sequenced so far. About 71% of these genes are homologous to *Arabidopsis* genes and 29% exist in cluster.

(2) Transposable elements. About one third of the rice genome is occupied by various kinds of transposons, retrotransposons and other families of transposable elements.

(3) Organelle insertions. About 0.2% of rice genome corresponds to sequences which are highly homologous to the chloroplast and mitochondrial DNA indicating recent transfer of these DNA fragments.

For the genome wide annotation, two automated

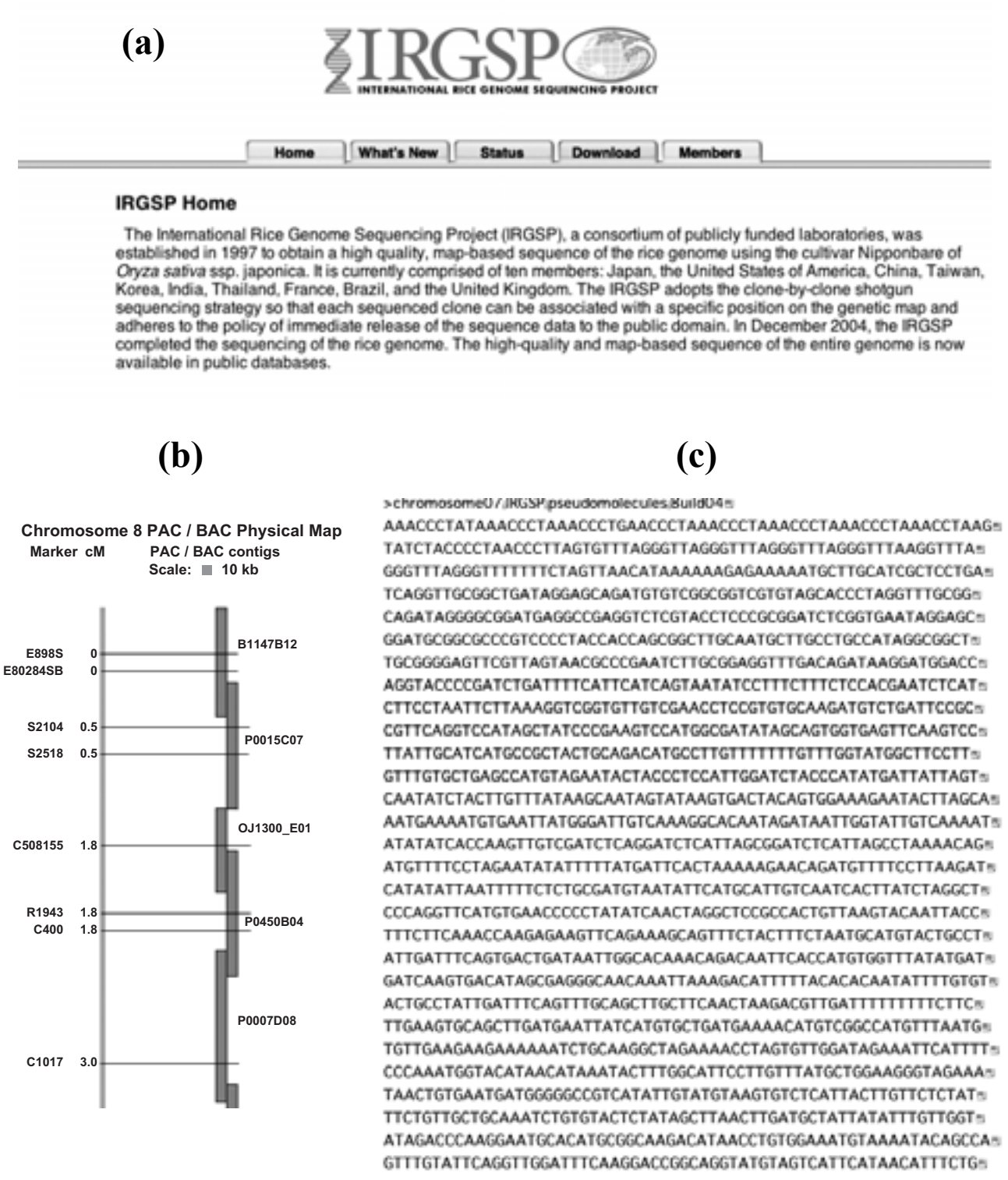

Fig. 2. Access to the rice genome sequence through the IRGSP website

(a): IRGSP homepage (front page)
(http://rgp.dna.affrc.go.jp/IRGSP/index.html).

(b): IRGSP physical map (part of chromosome 8).
Clone positions and corresponding DNA markers are shown.

(c): Part of the nucleotide sequence of rice chromosome 7
pseudomolecule (text file, downloaded).

genome annotation systems developed by RGP and TIGR (The Institute for Genomic Research), respectively, were used for analysis of the IRGSP genome sequence. The RiceGAAS (http://ricegaas.dna.affrc.go.jp/) from RGP automatically retrieves rice genomic sequences from the GenBank entries, analyzes the sequence with 5 kinds of gene prediction programs, and searches for homology with the rice ESTs, rice cDNAs, and known proteins in the databases. The annotation system developed by TIGR also facilitates comprehensive gene prediction and construction of gene models based on homology searches with annotation features enhanced by the TIGR PASA program[12]. An initial attempt to provide a standardized annotation of the rice genome sequence was organized through the First Rice Annotation Project Meeting (RAP1) held in Tsukuba in December 2004 and participated by plant biologists and informaticists from around the world. The curated gene models and database generated from this annotation jamboree will be published soon (Rice Annotation Project, submitted).

Functional genomics

Although many unique genes have been discovered in the rice genome, only 60% of the predicted genes are supported by ESTs or full-length cDNAs. On the other hand, about 40% of the predicted genes have no clues as to function (no match to the Swiss-plot entries). Although some of these unknown gene models may be false-positives, a significant number should be real since 71% of all the predicted genes have homologues in *Arabidopsis*. Recent analysis by tiling microarray designed for chromosome 4 indicates that 81% of the computationally predicted gene models have transcriptional activities[14]. Therefore elucidation of the function(s) of predicted genes is much more important. Forward genetics is the most powerful tool where genetic mapping could delimit a useful trait at the specific region of genome to identify the candidate gene. This method, positional cloning, also helps to map quantitative trait loci (QTLs) as Mendelian factors with the use of isogenic lines[25].

Recently, new genomic resources such as tagging lines in which specific genes are mutated artificially have been developed and put into usage. *Tos17*, one of the rice innate retrotransposons, has a tendency to be inserted in the genic region[20], establishing a series of the gene disruption lines that could be useful in identifying gene function and relating the genome with the phenome directly[23,31]. Other gene-tagging systems such as that using T-DNA are also utilized to clone the trait genes[15]. Regulated gene expression is the most important indicator of gene function. Monitoring the gene expression with oligomicroarray based on the sequence information from the comprehensive full-length cDNA collection[16] will let us know the global dynamics of rice gene expression. A 22K of rice oligomicroarray (G4138A) is now commercially available through Agilent Technologies.

Comparative genomics

Rice has been estimated to diverge from other cereal species such as barley, maize, and wheat from about 10–60 million years ago[2]. Although it is believed that domestication of rice occurred about 10,000 years ago, the present major subspecies, namely, *japonica* and *indica* must have diverged from a common rice ancestor at about 200,000 to 440,000 years ago[17,35]. The genus *Oryza* consists of 23 species adapted to a wide range of environmental conditions worldwide[34]. These diverse *Oryza* species could be a natural germplasm bank and rich sources of many novel genes which could be utilized for agriculture. Several research institutions including the International Rice Research Institute (IRRI), the National Institute of Genetics (NIG), and the Genebank at the National Institute of Agrobiological Sciences (NIAS) are making significant efforts in collecting, preserving, analyzing, and distributing materials and information on cultivated and wild rices. Recently, the Oryza Map Alignment Project (OMAP, http://www.omap.org/resources.html) has been organized in the US to develop materials for genetic analysis of wild rice species and to unravel the molecular biology of the genus *Oryza*. In Japan, the MAFF Genome Diversity Project aims to investigate the molecular mechanism of the reproductive barrier, which is a major obstacle to gene introduction across species.

The rice genome sequence is considered as a “gold standard” and is therefore pivotal to sequence comparison among the grass genomes. Several studies on microsynteny provide an overview on the magnitude of systenic regions among cereal crops[3,5,30]. These studies have shown that while colinearity is generally maintained at sequence level, some small insertions/deletions(indels), gene duplications and inversions may disrupt the microcolinearity[7]. To date, *Arabidopsis* and rice are the only plants for which the high-quality genome sequences have been revealed. These sequences will lead to a detailed comparison of monocot and dicot biosystems to verify how these systems are similar or how they differ. The high-percentage of the homolog detection between rice and *Arabidopsis* indicates that the basic system in plants may be the same, while some specific genes may have developed in cereal crops in the process of crop evo-

lution and domestication. Recent progress of DNA cloning and sequencing technology may pave the way for the initiation of genome sequencing in other crops (Whitelaw et al.[36], Wheat genome sequencing pilot project at IGROW, http://www.k-state.edu/igrow/IGROW_sequencing.html). Comprehensive genomic information of maize and wheat will be greatly useful for understanding how the crop genomes changed and evolved from a common ancestor.

Application to breeding

Various kinds of DNA markers such as RFLPs, SSRs, and SNPs have been constructed. Comparison of the Nipponbare sequence with the genome sequence of *indica* rice also shows a significant amount of SNPs and in-dels genomewide[9,13,39], which could be rich sources for polymorphic DNA markers throughout the genome. The DNA marker-assisted selection (MAS) is now widely applied in rice breeding as well as breeding of other crops. The established fine DNA markers at nucleotide level will identify the trait markers which are most proximal to the responsible gene, or dissect the neighboring genes to facilitate the introduction of only a desired trait by conventional breeding. The DNA marker-based breeding will be much more desirable than preexisting breeding strategies because selection of the introgressed lines can be performed even at a very early stage of development thereby reducing the time and labor required to establish new varieties.

Conclusion—Future aspects of rice biology and breeding

The completed rice genome sequence is revolutionizing rice biology. From now on, researchers will find it easier to identify corresponding genes for specific traits or for a specific response. Moreover, each gene function will be understood in the background of the gene network and traffic of the gene products. Integration of genomics with the newly emerging fields of proteomics, phenomics, metabolomics etc. will help us in understanding virtually all life processes as a dynamic system. An in-depth understanding of the rice genome structure will also pave the way for modification of metabolic processes to improve a specific agronomic trait without sacrificing other beneficial traits. The notion of "breeding by design" could be easily facilitated with the technical improvements in introducing novel genes, accumulating genes (pyramiding), and perturbing systems by modification of gene activities. This new concept of breeding will progress dramatically with the development of marker-assisted selection and genetic engineering. In this corollary, utilization of the genetic resources world wide including remote wild rice species as the main source of natural variation should be greatly promoted.

Acknowledgments

The authors thank all the members of the RGP and the IRGSP for their excellent work in rice genome sequencing. We also thank Dr. B. A. Antonio of NIAS for critical reading of the manuscript. This work is funded by the Ministry of Agriculture, Forestry and Fisheries (MAFF) of Japan (Rice Genome Projects, GS1101-1303).

References

1. Baba, T. et al. (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. Resour. Jpn.*, **14**, 41–52.
2. Bennetzen, J. L. & Ma, J. (2003) The genetic colinearity of rice and other cereals on the basis of genomic sequence analysis. *Curr. Opin. Plant Biol.*, **6**, 128–133.
3. Bowers, J. E. et al. (2005) Comparative physical mapping links conservation of microsynteny to chromosome structure and recombination in grasses. *Proc. Natl. Acad. Sci. USA*, **102**, 13206–13211.
4. Cheng, Z. et al. (2001) Toward a cytological characterization of the rice genome. *Genome Res.*, **11**, 2133–2141.
5. Delseny, M. (2004) Re-evaluating the relevance of ancestral shared synteny as a tool for crop improvement. *Curr. Opin. Plant Biol.*, **7**, 126–131.
6. Devos, K. M. & Gale, M. D. (1997) Comparative genetics in the grasses. *Plant Mol. Biol.*, **35**, 3–15.
7. Devos, K. M. (2005) Updating the 'Crop Circle'. *Curr. Opin. Plant Biol.*, **8**, 155–162.
8. Ewing, B. et al. (1998) Base-calling of automated sequencer traces using *phred*. I. Accuracy assessment. *Genome Res.*, **8**, 175–185.
9. Feltus, F. A. et al. (2004) An SNP resource for rice genetics and breeding based on subspecies *indica* and *japonica* genome alignments. *Genome Res.*, **14**, 1812–1819.
10. Feng, Q. et al. (2002) Sequence and analysis of rice chromosome 4. *Nature*, **420**, 316–320.
11. Fleischmann, R. D. et al. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, **269**, 496–512.
12. Haas, B. J. et al. (2003) Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.*, **31**, 5654–5666.
13. International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. *Nature*, **436**, 793–800.
14. Jiao, Y. et al. (2005) A tiling microarray expression analysis of rice chromosome 4 suggests a chromosome-level regulation of transcription. *Plant Cell*, **17**, 1641–1657.
15. Kang, H. G. et al. (2005) White-core endosperm *floury*

endosperm-4 in rice is generated by knockout mutations in the C_4-type pyruvate orthophosphate dikinase gene (*OsPPDKB*). *Plant J.*, **42**, 901–911.
16. Kikuchi, S. et al. (2003) Collection, mapping, and annotation of over 28,000 cDNA clones from *japonica* rice. *Science*, **301**, 376–379.
17. Ma, J. & Bennetzen, J. L. (2004) Rapid recent growth and divergence of rice nuclear genomes. *Proc. Natl. Acad. Sci. USA*, **101**, 12404–12410.
18. Mahairas, G. G. et al. (1999) Sequence-tagged connectors: A sequence approach to mapping and scanning the human genome. *Proc. Natl. Acad. Sci. USA*, **96**, 9739–9744.
19. Mao, L. et al., (2000) Rice transposable elements: A survey of 73,000 sequence-tagged-connectors. *Genome Res.*, **10**, 982–990.
20. Miyao, A. et al. (2003) Target site specificity of the *Tos17* retrotransposon shows a preference for insertion within genes and against insertion in retrotransposon-rich regions of the genome. *Plant Cell*, **15**, 1771–1780.
21. Moore, G. et al. (1995) Grasses, line up and from a circle. *Curr. Biol.*, **5**, 737–739.
22. Nagaki, K. et al. (2004) Sequencing of a rice centromere uncovers active genes. *Nat. Genet.*, **36**, 138–145.
23. Nonomura, K. et al. (2004) The novel gene *HOMOLOGOUS PAIRING ABERRATION IN RICE MEIOSIS1* of rice encodes a putative coiled-coil protein required for homologous chromosome pairing in meiosis. *Plant Cell*, **16**, 1008–1020.
24. Paez, J. G. et al. (2004) Genome coverage and sequence fidelity of phi29 polymerase-based multiple strand displacement whole genome amplification. *Nucleic Acids Res.*, **32**, e71.
25. Salvi, S. & Tuberosa, R. (2005) To clone or not to clone plant QTLs: present and future challenges. *Trends Plant Sci.*, **10**, 297–304.
26. Sasaki, T. (2003) Rice genome analysis: Understanding the genetic secrets of the rice plant. *Breed. Sci.*, **53**, 281–299.
27. Sasaki, T. et al. (2002) The genome sequence and structure of rice chromosome 1. *Nature*, **420**, 312–316.
28. Shevchenko, Y. et al. (2002) Systematic sequencing of cDNA clones using the transposon Tn5. *Nucleic Acids Res.*, **30**, 2469–2477.
29. Soderlund, C., Longden, I. & Mott, R. (1997) FPC: a system for building contigs from restriction fingerprinted clones. *CABIOS*, **13**, 523–535.
30. Sorrells, M. E. et al. (2003) Comparative DNA sequence analysis of wheat and rice genomes. *Genome Res.*, **13**, 1818–1827.
31. Tanaka, K. et al. (2003) Three distinct rice cellulose synthase catalytic subunit genes required for cellulose synthesis in the secondary wall. *Plant Physiol.*, **133**, 73–83.
32. The Rice Chromosome 10 Sequencing Consortium (2003) In-depth view of structure, activity, and evolution of rice chromosome 10. *Science*, **300**, 1566–1569.
33. The Rice Chromosome 3 Sequencing Consortium (2005) Sequence, annotation, and analysis of synteny between rice chromosome 3 and diverged grass species. *Genome Res.*, **15**, 1284–1291.
34. Vaughan, D. A., Morishima, H. & Kadowaki, K. (2003) Diversity in the *Oryza* genus. *Curr. Opin. Plant Biol.*, **6**, 139–146.
35. Vitte, C. et al. (2004) Genomic paleontology provides evidence for two distinct origins of Asian rice (*Oryza sativa* L). *Mol. Genet. Genomics*, **272**, 504–511.
36. Whitelaw, C. A. et al. (2003) Enrichment of gene-coding sequences in maize by genome filtration. *Science*, **302**, 2118–2120.
37. Wu, J. et al. (2003) Physical maps and recombination frequency of six rice chromosomes. *Plant J.*, **36**, 720–730.
38. Wu, J. et al. (2004) Composition and structure of the centromeric region of rice chromosome 8. *Plant Cell*, **16**, 967–976.
39. Yu, J. et al. (2005) The genomes of *Oryza sativa*: a history of duplications. *PLoS Biol.*, **3**, e38.
40. Zhang, Y. et al. (2004) Structural features of the rice chromosome 4 centromere. *Nucleic Acids Res.*, **32**, 2023–2030.