The Rice Genome Research Program (RGP) in Japan has been collaborating with the international community in elucidating a complete high-quality sequence of the rice genome. As the pioneer in large-scale analysis of the rice genome, the RGP has successfully established the fundamental tools for genome research such as a genetic map, a yeast artificial chromosome (YAC)-based physical map, a transcript map and a phage P1 artificial chromosome (PAC)/bacterial artificial chromosome (BAC) sequence-ready physical map, which serve as common resources for genome sequencing. Among the 12 rice chromosomes, the RGP is in charge of sequencing six chromosomes covering 52% of the 390 Mb total length of the genome. The contribution of the RGP to the realization of decoding the rice genome sequence with high accuracy and deciphering the genetic information in the genome will have a great impact in understanding the biology of the rice plant that provides a major food source for almost half of the world's population. A high-quality draft sequence (phase 2) was completed in December 2002. Since then, much of the finished quality sequence (phase 3) has become available in public databases. With the completion of sequencing in December 2004, it is expected that the genome sequence would facilitate innovative research in functional and applied genomics. A map-based genome sequence is indispensable for further improvement of current rice varieties and for development of novel varieties carrying agronomically important traits such as high yield potential and tolerance to both biotic and abiotic stresses. In addition to genome sequencing, various related projects have been initiated to generate valuable resources, which could serve as indispensable tools in clarifying the structure and function of the rice genome. These resources have been made available to the scientific community through the Rice Genome Resource Center (RGRC) of the National Institute of Agrobiological Sciences (NIAS) to enable rapid progress in research that will lead to thorough understanding of the rice plant. As the next trend in rice genome research will focus on determining the function of about 40,000–50,000 genes predicted in the genome as well as applying various genomics tools in rice breeding, an unlimited access to rice DNA and seed stocks will provide a broad community of scientists with the necessary materials for formulating new concepts, developing innovative research and making new scientific discoveries in rice genomics.

Keywords: Applied genomics — Database — DNA — functional genomics — Genome resources — Genome sequencing.

Abbreviations: BAC, bacterial artificial chromosome; CUGI, Clemson University Genomics Institute; EST, expressed sequence tag; IRGSP, International Rice Genome Sequencing Project; NIAS, National Institute of Agrobiological Sciences; PAC, phage P1 artificial chromosome; QTL, quantitative trait locus; RFLP, restriction fragment length polymorphism; RGP, Rice Genome Research Program; RGRC, Rice Genome Resource Center; RiceGAAS, Rice Genome Automated Annotation System; YAC, yeast artificial chromosome.

Introduction

The elucidation of the complete sequence of the genomes of many model organisms is probably the most significant scientific achievement in the last decade of the 20th century and the beginning of the 21st century. First and foremost, the genome sequence provides a bird’s eye view of the information needed for understanding the biology of model organisms. For rice, which is a major cereal crop that provides food to more than half of the world’s population, revealing the genome sequence is an achievement of great proportion considering the impact on what can be accomplished in other cereal crops such as maize, wheat, barley and sorghum (Sasaki 2003). Sequencing of the the rice genome was initiated in 1997 as an international collaboration among 10 countries with the common aim of attaining a finished sequence of the rice genome. The clone-by-clone sequencing strategy adopted by the sequencing consortium facilitated the immediate release of the sequence data to the public domain as soon as the sequence of a particular clone was completed. The road towards the completion of a high-quality draft sequence of the entire genome will be described here, with specific details on how the tools for genome analysis were developed and how they become useful not only for genome sequencing but also for other areas of rice research.

However, the genome sequence is just a framework or an outline that provides the essential information that will lead to
a thorough understanding of the physiology, morphology, genetics and development of cereal crops. The goals ahead for rice genomics will focus on understanding how the rice plant functions under normal conditions and how it responds to altered conditions based on the genome sequence, and ultimately using the information for rice improvement. Although analysis of genomic sequence data with gene prediction programs is a good starting point for gathering information about the genome, it is not sufficient to allow a true understanding of the final mRNA, protein or even regulatory elements encoded in the genome. The actual function of the gene can be determined only with sufficient information on mRNA structures, its disruption with the subsequent appearance of a phenotypic variant or from actual cloning of the genes underlying complex traits, and a thorough interpretation of the mechanism of gene expression. Thus, in addition to the rice genome sequencing, the National Institute of Agrobiological Sciences (NIAS) is currently pursuing various research projects encompassing different areas of structural, functional and applied genomics. In order to facilitate the management and distribution of biological materials generated from these projects, the Rice Genome Resource Center (RGRC) was established in 2003 (Antonio et al. 2003). The genome resources and informatics infrastructure generated from the rice genome project will also be described here, focusing on their significance in the future directions of rice research.

Genome Mapping

The Rice Genome Research Program (RGP) started the analysis of the rice genome in 1991 by constructing the fundamental tools for comprehensive genome analysis including genome sequencing. These tools and resources include a high-density linkage map, a yeast artificial chromosome (YAC)-based physical map, a transcript map and a sequence-ready phage P1 artificial chromosome (PAC)/bacterial artificial chromosome (BAC) physical map (Fig. 1).

Genetic map

A molecular linkage map is important for mapping phenotypic traits onto the genome. The RGP utilized the RFLP (restriction fragment length polymorphism) between Nipponbare (japonica) and Kasalath (indica) to generate a large number of DNA markers for linkage analysis. A single F2 population of 186 individuals derived from a cross between Nipponbare and Kasalath was used for genetic analysis, whereas cDNA clones from various tissues, organs and cultured cells as well as genomic clones were used as probes for Southern hybridization. The first genetic linkage map generated by the RGP consisted of 1,383 DNA markers (Kurata et al. 1994). With the mapping of additional markers, a more saturated genetic map with 2,275 DNA markers was established (Harushima et al. 1998). At present, the high-density linkage map consists of 3,267 markers [see Appendix 1 (1)]. These markers are useful not only for map-based cloning of agronomically important genes, but also as the source of polymorphic markers for genetic analysis of other rice varieties or other cereal crops. Moreover, these high-density linkage markers have been used in dissecting the quantitative agronomic traits into several genetic loci [quantitative trait loci (QTLs)], molecular cloning of mapped QTLs (Yano and Sasaki 1997) and marker-assisted selection (MAS) (Courtois et al. 2003, Yano et al. 2003) of the desirable genotypes for efficient breeding of desirable rice varieties.

YAC-based physical map

From the genome analysis based on the genetic linkage maps, the RGP proceeded with genome analysis based on molecular biology. Physical maps, in which genomic DNA of a particular region is cloned and aligned along the genome, are important resources for analysis of the structure of the genome. In human, both a YAC physical map (Wang et al. 1999) and a cosmid map (Soeda et al. 1995) were constructed. Considering the genome size of rice, the RGP constructed a physical map using a YAC library of 7,000 clones with an average insert length of 350 kb (Umehara et al. 1995). This library was screened with the RFLP markers by colony/Southern hybridization and the screened YAC clones were mapped onto the
positions of the genetic markers. The construction of the YAC physical map was performed in parallel with the construction of the genetic map. The genetic map with 1,383 markers generated a YAC-based physical map with 52% coverage corresponding to 222 Mb of the genome (Kurata et al. 1997) and the more saturated genetic map with 2,275 markers generated a YAC-based physical map with 63% coverage corresponding to 270 Mb of the genome (Saji et al. 2001). Although the YAC clones were not suitable as substrates for genome sequencing, the YAC-based physical map provided the framework that accelerated the positional cloning of agronomically important traits and the construction of a sequence-ready physical map.

Transcript map

To perform genome sequencing of rice, a sequence-ready physical map covering the entire genome with clones such as PAC and BAC is an absolute prerequisite. Considering the 100–200 kb average insert size of these clones, >4,000 anchors would be necessary to cover the entire rice genome. The RGP adopted a novel strategy of transcript mapping to increase the genomic anchor in a sequence-ready physical map. Paired polymerase chain reaction (PCR) primers were designed from the 3′-terminal sequences [expressed sequence tags (ESTs)] of unique cDNA clones and these 3′-ESTs were mapped by PCR screening with the physically mapped YAC clones as templates. The EST markers have high mapping efficiency due to site specificity of the 3′ sequence for PCR screening. In addition, polymorphism is not necessary to facilitate the mapping. The resulting 6,591 mapped EST markers within a rice transcript map are sufficient for positioning the PAC/BAC clones throughout the genome, except for the centromere and telomere region where the distribution of ESTs is scarce (Fig. 2, Wu et al. 2002). As these ESTs are part of the sequence from the expressed genes in rice, the EST map reveals directly the distribution of active genes, which is also helpful for studying the correlation between the genetically identified loci and the gene candidates expressed within the region. Moreover, the EST mapping has also contributed to both genetic and physical map construction. Part of the cDNAs used as EST markers were genetically analyzed and converted to linkage markers. Through genetic mapping, 431 ESTs were assigned to unmapped YACs, thereby increasing the genome coverage of the physical map. To date, the YAC-based physical map covers 81% of the rice genome. These EST markers have also been used as genetic markers for analysis of other rice varieties.

Genomic libraries/sequence-ready physical map

The RGP constructed two types of genomic libraries using PAC and BAC vectors to serve as template for the clone-by-clone sequencing strategy. A Nipponbare PAC library with 70,000 clones (Baba et al. 2000) and a BAC library with 50,000 clones (Wu et al. 2003) were used as the main resources for genome sequencing. The pooled PAC/BAC clones were screened by PCR using the PCR primers that were generated for mapping ESTs. Additional genome resources such as two BAC libraries (total of 90,000 clones) from CUGI (Clemson University Genomics Institute, Mao et al. 2000) and BAC clones with draft sequences donated by the Monsanto Co. to the International Rice Genome Sequencing Project (IRGSP) were used to increase map coverage. The CUGI BAC libraries had the additional information on fingerprint contigs and the end sequences of each BAC clone. The RGP is in charge of six of the 12 chromosomes, namely chromosomes 1, 2, 6, 7, 8 and 9, for genome sequencing (Fig. 3, Sasaki 2004). The sequence-ready physical maps for these chromosomes were constructed by aligning PAC/BAC clones using EST markers and localizing the fingerprint BAC contigs in case a member clone was already mapped by markers. Many gaps between the contigs were filled by the STC method, end-walking strategies and utilization of a fosmid library. Currently, the sequence-ready physical map covering chromosome 1 of rice is shown in Fig. 2 (left).
physical maps for the six chromosomes assigned to the RGP are covered by PAC/BAC contigs from about 1,800 clones corresponding to nearly 190 Mb of the six rice chromosomes or about 99% of the euchromatic regions [see Appendix 1 (2)], indicating that the physical maps for these chromosomes are almost completed. An example of the sequence-ready map with aligned PAC/BAC clones from various sources is shown in Fig. 4. Only the centromere of chromosome 8 has been covered by PAC/BAC clones, suggesting that the centromeric regions of other chromosomes may be rich in repeat sequences that hampered the identification of clones.

Genome Sequencing

High-quality genome sequencing is the most important step for acquiring accurate information on rice so that it can be used as a gold standard for analysis of crop plants. The RGP as well as the international consortium set a 99.99% sequence accuracy based on the Bermuda standard [see Appendix 1 (3)].

RGP sequencing initiative

The RGP has divided the whole step of sequencing into two parts. The first half is a streamlined process for draft sequencing, and the second half is a manual process for finishing, annotation and data submission. The PAC/BAC clones, which comprise the minimum tiling path of the physical maps, were used for sequencing. The clones were cultured, the insert DNAs were purified, and shotgun libraries with insert sizes of 2 and 5 kb were constructed using a plasmid vector. A total of 2,000 subclones were used for sequencing from both ends of the inserted DNA by capillary sequencers. These draft sequences (a total of 4,000 sequences, over 10 times coverage of the clone size) with a quality (error probability) for every nucleotide were automatically collected, and assembled by Phred/Phrap software (Ewing and Green 1998) to construct the sequence contigs. Among the PAC/BAC clones with the assembled contigs, only those clones in which all the contigs are ordered and oriented with a small number of gaps (GenBank HTG phase 2) were released to the public domain through the DDBJ. Although these draft sequences have gaps within clones, the 10 times coverage indicates that these sequences are relatively accurate. The draft sequences are then completed (‘finishing’ step) as follows: (i) filling of sequence gaps by sequencing a bridge subclone over a gap; (ii) improving the quality by re-sequencing for ambiguous nucleotides; and (iii) resolving the misassembled contigs for the right position. This process facilitates the production of a contiguous sequence with 99.99% accuracy calculated by the summation of the quality of each nucleotide. These completed sequences are again submitted to DDBJ to update the previous submission.

The PAC/BAC sequences are then annotated through several routine processes. The gene predictions by programs such as Genescan (Burge and Karlin 1997), FGENESH [see Appendix 1 (4)] and Genemark [see Appendix 1 (5)], BLAST (Altschul et al. 1990) homology search against a protein sequence database (nr), and the mapping results of rice ESTs [see Appendix 1 (6)].
From mapping to sequencing, post-sequencing and beyond

dix 1 (6) and full-length cDNAs (Rice Full-length cDNA Consortium 2003) onto the PAC/BAC sequences are used for constructing gene models. The resulting gene models are manually curated using an in-house annotation editor (Annotation Plot) to construct the most plausible structure of the predicted gene based on existing evidence. The gene models as well as the suggested protein functions are submitted to DDBJ and added to the sequence. The RGP published the overall analysis of chromosome 1 (Sasaki et al. 2002), which is the largest chromosome in rice. The other chromosomes that have been completely sequenced are chromosome 4 (Feng et al. 2002) and chromosome 10 (Rice Chromosome 10 Sequencing Consortium 2003). The rice genome is characterized by a high GC content in the exon domain and frequent tandem gene duplications. It also has many kinds of transposable elements (transposons, retrotransposons, MITEs). As of October 2004, almost all of the PAC/BAC clones on the RGP physical maps have been completed. The remaining clones consisted of the most difficult clones for sequencing because of the highly repeated sequences. More than 1,700 clones have been manually annotated and submitted to databases after sequence completion [see Appendix 1 (7)].

Analysis of the centromeric region

The centromere is one of the most important structures in the eukaryotic chromosomes. It is the site for kinetochoore complex formation and also a key factor for the accurate chromosome distribution to daughter cells during cell division. Functional analysis of the centromere was done mostly in yeast chromosome. The complex nature of the chromosomes of higher eukaryotes has made it difficult to study the detailed centromere structure and function. Physical contigs were constructed in the centromeric region of chromosome 8. Previous cytological analysis revealed that this chromosome has the shortest array of the CentO repeat, which makes it relatively easy to analyze (Cheng et al. 2002). Thus the RGP has succeeded in constructing a PAC/BAC contig spanning the centromeric region of chromosome 8. These PAC/BAC clones were subjected to genome sequencing. Several techniques for sequence finishing have completed the 2 Mb sequences containing the CentO repeat (Wu et al. 2004). The core region of the centromere has 68.5 kb of 155 bp CentO repeats which were divided into three clusters. More than 220 transposable element-related sequences (most of which belong to the RIRE family) were also found in the same region. Surprisingly there were also protein genes that may be functional in rice cells (some expressed genes were mapped here). Independent analysis by another group revealed virtually the same complicated nature of this centromeric region (Nagaki et al. 2004). The centromeric regions of chromosomes 4 and 5 have also been covered by BAC contigs. Comparing the sequences with the centromeric sequence of rice chromosome 4 in which CentO sequences are dispersed (Zhang et al. 2004), it is assumed that the centromere of each chromosome has a unique structure in rice. There is a large variation of CentO contents in each chromosome. It might be intriguing to speculate the evolution of rice chromosomes by comparing the structure of centromeres. Among the divided CentO clusters, a major cluster (223 units) was cloned in one BAC clone while the other two (50 and 169 units, respectively) were on the other BAC. This separate cloning of major CentO clusters made it possible to identify the exact positions of 442 of almost exactly the same tandem CentO repeat sequences, thus showing the advantage of the clone-by-clone strategy.

International Sequencing Collaboration

The international rice community proposed the sequencing of the rice genome in a workshop held in Singapore in Sep-

---

**Fig. 4** A detailed physical map of the 102.9–107.7 cM region of chromosome 8. The YAC-based physical map (green bar) is shown with the genetic markers, genetic distance and all aligned PAC/BAC clones. The PCR markers (STS and EST) used for screening are indicated with the relative position in the PAC or BAC clone. Five RGP-PACs (red), one RGP-BAC (brown), one CUGI-BAC (blue) and six Monsanto-BACs (yellow) were aligned in this region of chromosome 8. After checking the overlaps, some clones (marked by stripes) were found to be redundant and were not used for further analysis.
dix 1 (8)]. A total of 3,380 PAC/BAC clones with a total length sequence of the entire genome on December 2002 [see Appendix A]. The IRGSP successfully completed a high-quality draft sequence by the early 1999, using the Monsanto BACs with 5× coverage. IRGSP has accelerated the rate of finishing the rice genome (Venter et al. 2001). The IRGSP had consistently established as an international collaboration with the ultimate goal of sequencing the entire rice genome. Currently, the IRGSP consists of publicly funded sequencing groups from 10 countries and regions. The IRGSP adopted a clone-by-clone sequencing strategy, chromosome sharing policy, and set a guideline for the published sequence accuracy and the immediate data submission to the public database. The RGP provided the genetic map information of the 12 chromosomes. Two PAC and BAC libraries from RGP and two CUGI BAC libraries as well as the information of the fingerprint contigs and BAC end sequences were utilized for the construction of the sequence-ready physical map.

Genome sequencing was also pursued by private companies with interest in more practical applications of the genome sequence information. The Monsanto Co. released a draft sequence of the rice genome obtained by a clone-by-clone strategy in 2000. These draft sequences were eventually donated to IRGSP. On the other hand, Syngenta-Myriad used the whole-genome shotgun sequencing which was first adopted for assembly of the bacterial genome, but later introduced into the Drosophila genome (Myers et al. 2000), mouse genome (Mouse Genome Sequencing Consortium 2002) and human genome (Venter et al. 2001). The IRGSP has consistently argued that the clone-by-clone method was the best strategy to obtain a high-quality finished sequence so that the rice genome sequence can be a gold standard for monocots. However, with the announcement of the ‘completion’ of sequencing by the private sectors, there emerged a strong desire for the publication of the entire sequence with higher quality and positional information even if it is not completed. From the year 2000, the IRGSP has accelerated its publication of the phase 2 ‘high-quality draft’ sequence to the public database. It effectively used the Monsanto BACs with 5× shotgun sequences and constructed a sequencing pipeline to accelerate the sequencing. As a result, the IRGSP successfully completed a high-quality draft sequence of the entire genome on December 2002 [see Appendix I (8)]. A total of 3,380 PAC/BAC clones with a total length of about 460 Mb were submitted. The RGP contributed 55% of the total sequence production. The sequence length excluding overlaps is 366 Mb, corresponding to about 92% of the rice genome. The IRGSP celebrated this accomplishment in a commemorative ceremony held in Tokyo in December 18, 2002 with the attendance of IRGSP members as well as researchers and government officials from Japan and various countries. Then the IRGSP set the next and the final goal of finishing the genome to phase 3 (finished quality) level on December 2004. This was a tough timeline for IRGSP because at that time >2,000 PAC/BAC clones were still at phase 2 quality. The RGP, with almost half of the total clones to be completed, constructed a finishing pipeline in which the trained finishers could effectively fill the gaps, improve the low-quality sequences and resolve discrepancies. New strategies for accurate sequencing were incorporated in the finishing pipeline including transposon sequencing, modification of DNA polymerase reaction conditions for efficient sequencing of GC- or AT-rich sequences, disruption of a higher order structure in the template by sonication, and the utilization of new capillary sequencers to acquire the stronger signal for basecall and longer reads. Furthermore, the IRGSP held a finishing workshop in Tsukuba, Japan in February 2004 to provide a forum for exchange of strategies among finishers from the participating countries. As a result, the IRGSP has accelerated the rate of finishing to as much as 143 clones per month from the previous rate of 95 clones per month (as of December 2003). At the time of writing (October 2004), 98% of the 3,453 PAC/BAC/fosmid clones in the mini-path of the physical map have been completed and submitted to public databases. In addition, several clones have been mapped successfully in the centromeric and the heterochromatic regions. Sequencing of these clones revealed various kinds of repeat sequences, which hindered accurate analysis. Although these clones have not been completely sequenced, the type and the composition of these repeats have been fully characterized. Along with the sequence completion, the size of gaps between contigs is analyzed by fiber-fluorescence in situ hybridization (FISH) analysis to determine the exact sizes of each rice chromosome, and accordingly the total size of the rice (Nipponbare) genome. The sequenced region is 370 Mb, which would be >95% of the entire genome (390 Mb is used as the tentative genome size). If this assumption is correct, the rice genome sequencing project conducted by IRGSP is almost near completion.

Tools for Map-based Genomics

Much of the post-genome research will depend on a robust informatics infrastructure for analysis and maintenance of genome sequence information. In order to extract as much biological information from the rice genome sequence as possible and to facilitate its utilization in understanding the genome structure of other biological systems, the RGP has developed an informatics infrastructure aimed at analyzing the genome sequence, developing integrated databases and releasing the analyzed sequence through the Internet (Fig. 5).
RiceGAAS: Rice Genome Automated Annotation System

RiceGAAS is a rice genome automated annotation system. This system integrates programs for prediction and analysis of protein-coding gene structure.

Integrated software is coding region prediction programs (GENSCAN, RiceHYM, FGENSEH, MEG), splice site prediction programs (SplicePredictor), homology search analysis programs (Blast, HHSEARCH, PROFILESEAR, MOTIF), RNA gene prediction programs (TRNAscan-SE), repetitive DNA analysis programs (RepeatMasker, Printrepeats), signal scan search program (Signal Scan), protein localization site prediction program (PSORT), and program of classification and secondary structure prediction of membrane proteins (SOSUI).

Blast against full-length cDNA sequences of japonica rice is integrated. The full-length rice cDNA sequence is provided by ROME database.

Interpretation of the coding region is fully automated and gene prediction is accomplished without manual evaluation and modification. Therefore some differences exist between the predicted genes by the system and the manually predicted genes included in the GenBank entries. At present about 84% of auto and manually predicted genes are the same at nucleotide level (see "comparison table of gene prediction").

Rice Genome Annotation Database (RAD)

In an effort to manage and integrate accumulated information efficiently from annotation of the genome sequence, we have also developed a database for manually curated gene models. The Rice Genome Annotation Database [see Appendix 1 (10)] system allows merging of the annotation of individual PAC/BAC clones and provides a graphical view of the genome sequence with relevant annotation information at contig.
tig level. It also facilitates gene search, statistical analysis of the characteristic features of predicted genes and efficient management of the genome annotation. The gene prediction data of a PAC/BAC clone together with the clone sequence are seamlessly united into a contig without redundant sequences and genes. A graphical view of individual genes in the context of the large sequence contigs can be particularly useful in searching for genes within a region of interest. It also provides statistical analysis for various features of the sequence such as average gene size, GC content, exons/introns, etc. Most of the annotation data for the six chromosomes (1, 2, 6, 7, 8 and 9) are already incorporated in RAD. For the rest of the chromosomes, GenBank files with annotation are subsequently retrieved, and incorporated into RAD to construct a comprehensive, non-overlapping database of manually curated genes from the rice genome.

INE [INtegrated Rice Genome Explorer; see Appendix 1 (11)] is a rice genome database for map-based genomics (Sakata et al. 2000). It has an integrated genome browser, which surveys all maps of the RGP: genetic map, YAC-based physical map, transcript map and PAC/BAC maps for the presentation of the genome sequence and the annotation. This database integrates the genetic information with a marker position as a key factor within the same browser; it goes across from the genetics world to molecular biology, and vice versa. The integrated maps for each chromosome provide a general overview of the genomic information such as the distribution of DNA markers, the position of ordered YAC clones and the sequence-ready physical map with PAC/BAC contigs. Several chromosomes can be displayed simultaneously for direct comparisons between or among chromosomes. The DNA markers in the genetic map can be traced to the physical map, transcript map and the genome sequence. Links are also provided for mapping information such as Southern hybridization data, image sets, sequences, as well as detailed contig maps in the case of sequenced regions of the chromosome. The annotation map provides details of the structural features of the predicted genes as well the results of homology searches with non-redundant protein and EST databases.

These databases and tools would be very beneficial to the understanding of the rice genome. In addition, they can provide a platform for map-based cloning, marker-based genomics and comparative genome analysis among cereal crops.

Access to Genome Resources

In addition to the rice genome sequencing being undertaken in collaboration with the IRGSP, the NIAS is currently pursuing various research projects encompassing different areas of rice genomics. Among the major projects are the rice full-length cDNA project, the Tos17 mutant panel project, functional analysis of genes relevant to agronomically important traits, comparative genomics of plant species assisted by rice genome information, development of rice genome simulators, elucidation of functions of useful genes by gene expression monitoring systems, and identification and analysis of proteins for gene discovery. The RGRC is in charge of maintenance and distribution of the biological materials generated from these projects. The genome resources currently being distributed by the RGRC [see Appendix 1 (12)] include the rice full-length cDNA clones, Tos17 mutant lines and plant materials for genetic analysis (Fig. 6).

Rice full-length cDNA

Identification of genes from the sequence data is greatly facilitated by full-length cDNA information. Such information is important for the confirmation of genes identified by genome annotation using various prediction programs and for identifi-
cification of alternative splicing sites for RNAs. The Rice Full-length cDNA Project has generated sequence data on 175,642 rice full-length cDNAs clustered into 28,469 non-redundant clones (Rice Full-length cDNA Consortium 2003). The starting materials for library construction represent various tissues such as root, shoot and panicle as well as calli subjected to various treatment and growth conditions, and therefore represent the complete array of genes expressed in rice. A total of 21,596 clones have been assigned with tentative protein functions through homology searches of publicly available sequence data. In addition, >94% of the clones could be mapped to subspecies japonica and indica genomic sequences, indicating that the nucleotide sequences of the gene-coding regions are very similar in these two subspecies. All sequence data are available through the database KOME [see Appendix (13)]. The database provides the nucleotide sequence and encoded amino acid sequence information, results of the homology search with the public databases, mapping information, the pattern of alternative splicing, protein domain information, transmembrane structure, cellular localization and gene function by gene ontology. Access to specific information for each full-length cDNA clone can be made by BLAST search, accession number of the clone, specific domain name and general key word search. The availability of rice full-length cDNA clones will greatly enhance the rate of gene discovery, patterns of splicing and the understanding of gene function and protein interactions in rice. Since the annotation of the sequence simply classifies various regions of the genome into genic regions, non-transcribed region, transposable elements, etc., homology with full-length cDNA sequence will address whether the predicted genes are ever expressed or not in order to facilitate more informative characterization of the gene.

**Tos17 mutant lines**

An alternative to the nucleic acid characterization through full-length cDNA homology is by direct demonstration that the gene has a function. This can be done by disrupting the gene function via inserting either a T-DNA or a transposon sequence. The insertional mutagenesis approach is expected to play an important role in determining the function of the genes predicted in rice. A transposon tagging strategy, which utilizes a rice endogenous retrotransposon Tos17, has been found to be effective in functional characterization of rice genes (Miyao et al. 2003). Mutations due to Tos17 transposition are normally induced under tissue culture conditions and are inherited in subsequent generations to facilitate analysis of the mutated gene. So far, a Tos17 mutant panel with >50,000 insertional mutant lines carrying about 500,000 insertions has been generated. These resources would be very useful for forward and reverse genetics analyses. The Rice Tos17 Insertion Mutant Database [see Appendix (14)] currently contains flanking sequences of Tos17 insertion sites from 5,000 lines and associated phenotypes. By performing a simple BLAST search against these flanking sequences, any gene of interest can be easily found if the collection contains a mutation in that gene. The disruption of gene function by insertional mutagenesis is generally applicable for rice because it can be easily transformed and regenerated. The insertion of Tos17 into the coding region can induce partial or complete inactivation of the gene that may therefore lead to identification of the function of the gene.

**Genetic analysis materials**

Analyses of agronomically useful genes and complex traits including QTLs have been effectively accomplished using a number of resources including a high-density linkage map, appropriate crosses and subsequent generations of segregating populations to allow the mapping of the traits of interest. A strategy that involved developing specific mapping populations and fine mapping of many traits proved effective in map-based cloning. The genetic populations currently available include Nipponbare/Kasalath Backcross Inbred Lines (BIL); Nipponbare/Kasalath Chromosome Segment Substitution Lines (CSSL); Koshihikari/Kasalath Backcross Inbred Lines (BIL); Koshihikari/Kasalath Chromosome Segment Substitution Lines (CSSL); Sasanishiki/ Habataki Backcross Inbred Lines (BIL); and Akihikari/Koshikihari Doubled-haploid Lines (DHL). These materials have been successfully used in identifying many QTLs for agronomically important traits (Yano et al. 2000, Yano et al. 2003). However there are many key agronomic characteristics that remain to be analyzed in rice. In addition to a wide array of molecular markers for rice, the availability of genetic populations with well-characterized chromosomal segments will be very useful for fine mapping of QTLs involved in flowering time, seed development, disease resistance, etc. in the context of germplasm diversity.

**Rice Genomics in the Post-sequencing Era**

The map-based, accurate sequence of Nipponbare sequence will soon be a gold standard for studying genes for phenotypic traits in other varieties in japonica species, thus contributing to the map-based cloning and QTL analysis of agronomic traits, and the acceleration of rice breeding of the desired traits. The indica species is the most cultivated rice species in the world. The indica genome and genes are also genetically studied through several japonica–indica crosses. The comparative study between japonica and indica species reveals that although the basic structure is conserved, there are many base substitutions and insertions/deletions between these species. The Nipponbare sequence will also be beneficial to understand indica genomes. The RGP has also succeeded in in silico mapping of BAC end sequences from the Kasalath variety, one of the parental varieties of rice genetic maps. This in silico map will be useful in determining the corresponding sequences in the desired regions of the indica genome. The complete rice sequence will in turn be invaluable information to study genomics of the grasses. Recently, the maize genome project that aims to reveal all the gene-rich regions has been launched.
(Martienssen et al. 2004). One of the coverage checks of the gene domains for filtered genome sequence is to compare them with the rice genome. Since many other grass genomes would be analyzed only for the gene-rich regions, the high-quality rice genome sequence will be a good standard for evaluating the gene coverage. Moreover, the complete rice sequence will be useful in developing genetic markers of the desired region in the grass genomes, thereby delimiting the region of interest particularly in genomes much larger than rice.

The increased understanding of the structure and function of the rice genome and the ability to manipulate those functions will have far-reaching implications in the future of rice research. Many of the discoveries concerning the function of particular genes that will be elucidated from forward and reverse genetic methods can be included in conventional breeding techniques. Since only a few plant species are being sequenced, the availability of a complete rice genome sequence will be incredibly useful for comparative genomics. This should start with the cereal genomes by utilizing the syntenic relationships of the major cereal crops. Comparisons across species will be useful in understanding the basis of major biological processes including the complex metabolic pathways that control growth and development. These studies will require advanced genetic tool kits and resources that can be used to formulate new hypotheses. It will be essential from now on to coordinate the data produced from extensive genome analysis, the biological resources to provide evidence to gene expression and the emerging technologies in genome informatics. This will stimulate rice genomics research at various levels and by a wide variety of specialists including molecular biologists, plant physiologists, breeders and informaticists. It is also essential to develop a network that will link all fundamental rice genomic resources, databases such as the huge amount of data from rice proteome analysis, full-length cDNA, micro-array, insertional mutagenesis and biochemical profiling; and simulation of various parameters. In the long run, the desired success in rice genomics will involve developing the necessary tools for analyzing biological and agricultural problems in multiple dimensions.

Appendix 1

(2) http://rgp.dna.affrc.go.jp/cgi-bin/statusdb/irsgp-status.cgi
(4) http://www.softberry.com/berry.phtml
(5) http://opal.biology.gatech.edu/GeneMark/
(6) http://bank.dna.affrc.go.jp%/7expri/hiho/
(7) http://rgp.dna.affrc.go.jp/cgi-bin/statusdb/status.pl
(8) http://rgp.dna.affrc.go.jp/rgp/Dec18_NEWS.html
(9) http://ricegaas.dna.affrc.go.jp
(10) http://rgp.dna.affrc.go.jp/giot/INE.html
(11) http://rad.dna.affrc.go.jp
(12) http://www.rgc.dna.affrc.go.jp
(13) http://cdna01.dna.affrc.go.jp/cDNA/
(14) http://tos.nias.affrc.go.jp

Acknowledgments

The authors thank all the members of RGP, both past and present, for their dedicated work in rice genome sequencing. This work is funded by the Ministry of Agriculture, Forestry and Fisheries of Japan (Rice Genome Projects GS1101, 1102, 1103, 1201, 1301, 1302 and 1303).

References


Downloaded from https://academic.oup.com/pcp/article-abstract/46/1/3/1815112 by guest on 05 April 2019

(Received October 28, 2004; Accepted November 10, 2004)