

Perspective

Rice 2020: A Call For An International Coordinated Effort In Rice Functional Genomics

Qifa Zhang^{a,1}, Jiayang Li^b, Yongbiao Xue^b, Bin Han^c and Xing Wang Deng^{d,e}

a National Key Laboratory of Crop Genetic Improvement, National Center for Plant Gene Research (Wuhan), Huazhong Agricultural University, Wuhan 430070, China

b Institute of Genetics and Developmental Biology, Chinese Academy of Sciences and National Center for Plant Gene Research, Beijing 100101, China

c National Center for Gene Research & Institute of Plant Physiology and Ecology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, 500 Caobao Road, Shanghai 200233, China

d National Institute of Biological Sciences, Zhongguancun Life Science Park, Beijing 102206, China

e Department of Molecular, Cellular and Developmental Biology, Yale University, New Haven, CT 06520–8104, USA

ABSTRACT We describe a call for an international coordinated effort in rice functional genomics in the form of a project named RICE2020. The mission of the project will be: to determine the function of every gene in the rice genome by the year 2020, to identify functional diversity of alleles for agriculturally useful genes from the primary gene pool of rice, and to apply the findings of functional genomics research to rice genetic improvement.

INTRODUCTION

Rice is the main staple food for a large segment of the world population. In the last half-century, rice yield has more than doubled in most parts of the world and even tripled in certain countries within a period of four decades, from the 1960s to the 1990s (<http://faostat.fao.org/faostat/collections?subset=agriculture>), primarily as the result of genetic improvement. Thus, increasing rice production has been considered as an effective strategy for increasing global food production and safeguarding food security.

Rice has also become a model for the genomic study of monocotyledon species, due its small genome size, availability of the whole genome sequences of both *indica* and *japonica* subspecies, highly efficient transformation technology, and abundant genomic resources. The finished quality sequence based on the *japonica* cultivar Nipponbare shows that the rice genome is 389 Mb (International Rice Genome Sequencing Project, 2005) and encodes ~32 000 genes (The Rice Annotation Project, 2007, 2008), both of which are smaller than previous estimates. It is anticipated that the number of genes will change with the rapid evolving definition of genes with newly gained knowledge in genomic research.

Despite the tremendous progress made by rice researchers, there is still a huge gap of knowledge for bridging the genotype and phenotype, which is essential for breeding elite varieties suitable for sustainable agriculture. To this end, a highly coordinated effort that brings together scientists and resources worldwide is a desirable choice and perhaps the only practical and efficient one. We thus propose an Inter-

national Rice Functional Genomics Project (IRFGP), with a goal to determine the function of every gene in the rice genome by the year 2020, to identify functional diversity of alleles for agriculturally useful genes from the primary gene pool, and to apply the findings of functional genomics research to rice crop genetic improvement and beyond.

AN OUTLINE OF SCIENTIFIC OBJECTIVES

We propose the following objectives for this international effort, with elaboration of specific aims to be achieved.

(1) Development of Enabling Tools and Genetic Resources for an International Community of Scientists to Conduct Functional Genomics Research in Rice

Under this objective, we propose three main aims to be achieved. Those are: (1) insertion mutant collections, (2) full-length cDNA collections, and (3) artificial micro-RNA (amiRNA) collections.

Large numbers of insertional mutant lines have been generated globally using mainly T-DNA and transposable elements

¹ To whom correspondence should be addressed. Email qifazh@mail.hzau.edu.cn.

© The Author 2008. Published by the Molecular Plant Shanghai Editorial Office in association with Oxford University Press on behalf of CSPP and IPPE, SIBS, CAS.

doi: 10.1093/mp/ssn043, Advance Access publication 13 August 2008

Received 19 June 2008; accepted 20 June 2008

(Hirochika et al., 2004). Based on the current annotation of 32 000 genes and the estimated size 389 Mb of the rice genome, and assuming the random insertion sites in the genome, a total of 587 345 independent insertions are needed to obtain at least one insertion per gene with a probability of 0.99. With an average of about two copies of T-DNA insertions per line, as reported currently for the T-DNA insertion libraries (Jeon et al., 2000; Wu et al., 2003), around 300 000 independent transformants would be needed to saturate the genome. Although the number of T-DNA insertion transformants accumulated over the years is already larger than this number, the flanking sequence tags (FSTs) isolated so far are far short of this expectation. Thus, the current goal here is to generate FSTs from the sufficiently large number of T-DNA insertion lines and to make those lines with known FST available to the international community without any restriction. Collections of mutants generated with other technologies may also provide an important complement for regions and genes that are unable to be mutagenized by T-DNA insertions.

A large and complete full-length cDNA collection for all estimated 32 000 rice genes will be an important resource that can facilitate many research projects because it would save considerable workload by individual researchers. Such a collection, once comprehensive enough, will also improve the status of rice as a model cereal species. Full-length cDNAs from both subspecies *indica* and *japonica* cultivars have been isolated, with the number of independent clones totaling over 50 000 (The Rice Full-Length cDNA Consortium, 2003; Xie et al., 2005; Liu et al., 2007). However, the coverage of the available cDNA clones only reached 60% of estimated rice genes at best; thus, it is worthwhile to further continue this effort to reach a near saturation of the rice genome for a full-length cDNA for every rice gene. In addition, analyses with reference to the Nipponbare genome sequence showed that there are *indica*-specific cDNAs or vice versa, indicating that isolation of full-length cDNAs from both *indica* and *japonica* should be worth continuing for completeness of the full-length cDNA effort.

Recent studies have demonstrated in both *Arabidopsis* and rice that amiRNA can be a highly specific gene silencing tool (Schwab et al., 2006; Warthmann et al., 2008). It is now feasible to develop an amiRNA collection, with one amiRNA for each rice gene, and produce individual amiRNA transgenic rice line collections for phenotype characterization and distribution. This resource will nicely complement the insertion mutant collection, as it provides normally a partial loss-of-function phenotype and can be developed for silencing a small gene family at the same time, as well as non-homologous multiple genes with a single construct.

(2) Assignment of Biological Functions to Every Annotated Gene

We propose two aims for this objective: (1) systematic phenotyping and characterization of the mutants, and (2) systematic characterization of gene families.

A major task for characterizing the function of the genes is to examine possible phenotypic changes associated with the genes' loss-of-function mutations. For rice as a crop, it needs high-yielding, superior-quality, multiple resistances to biotic and abiotic stresses, and high nutrient use efficiency, among other characteristics. Many of the phenotypic changes can be observed under normal growth conditions, while other changes are conditional or inducible, which can be observed only under certain conditions. Moreover, detection of many traits, such as endosperm composition and nutrient efficiency, may require chemical and/or physical analyses, some of which may need sophisticated facilities. All this suggests the need for broadening the range of trait examination and enhancing the phenotyping capacity. Thus, in conducting phenotyping, the rice plants should be planted in multiple growing conditions, including biotic and abiotic stresses, and soils with low nutrients, and observed and examined using a variety of techniques.

It is essential to follow systematic approaches in order to have all the known genes included in the analysis, to relate the gene to phenotypic changes, thus eventually assigning a defined biological function to every gene. In achieving a systematic approach, a non-redundant set of mutant lines representing all the annotated genes should be identified and subjected to phenotyping experiments. The availability of FSTs of the mutant stocks and genome sequence would allow the tasks to be divided by chromosomes, by gene families or by some other classification among the participating groups, to ensure the completeness of the genome coverage. The amiRNA transgenic line collection should also be used as an important complementing approach. For those rice genes without T-DNA insertion mutants or with lethal phenotype, amiRNA transgenic lines should provide information that could not be obtained from insertion mutant lines alone. In the course of IRFGP, the developments and expertise in *Arabidopsis* functional genomic research should be closely followed.

(3) Systems-Wide Epigenomes, Gene Expression Profiles and Regulatory Networks

This objective is proposed to include three aims to be achieved. They are: (1) comprehensive cell- or tissue-specific epigenomes and transcriptomes for selected developmental stages, abiotic and/or biotic conditions; (2) identification of regulatory elements based on the epigenetic profiles and transcriptomes; and (3) systematic characterization of regulatory hierarchy of genome expression, its relationship to epigenomes during development and responses to various environmental changes, and their effects on growth and development.

Epigenetic modification states and gene expression profiles provide important clues about the function of the gene, while recent advances in ultra-throughput sequence technology provide an accurate and efficient means for generating cell- or tissue-specific epigenomes and transcriptomes. It is thus feasible to obtain accurate expression profiles of all the annotated genes and all selected epigenetic modification patterns during the entire lifecycle of the rice plants on a set of cell types

and tissues under specific treatments with very broad coverage. Expression profiles and epigenetic modifications, especially those cell-specific ones and ones responsive to changing environments, also provide data for identifying promoters and *cis*-elements regulating temporal and spatial expression. It is also important that multiple genotypes, or probably a core set of genotypes, should be used in the profiling analysis for possible expression polymorphism and epigenomes. An important issue is to apply a technology that can select cell types or key tissue types for transcriptome analysis, although the recent advance of laser cell capture technology should make this objective feasible. Once comprehensive collections of transcriptomes and epigenomes are generated for rice, it should permit informatic analysis of the regulatory network for gene expression and identification of *cis*-regulatory elements.

(4) Global Analyses of the Proteome and Protein–Protein Interactions

We propose two main aims for this objective: (1) tissue-specific proteomes of selected developmental stages and under selected defense and stress conditions; and (2) an experimentally defining comprehensive protein–protein interaction network.

Most of the biological functions are performed by proteins, and frequently changes in mRNA levels are not directly related to the quantity and activity of the proteins. Understanding protein dynamics will enable prediction of the functional machinery working throughout a plant's lifecycle. A combination of two-dimensional electrophoresis and mass-spectrometry may be used to identify cellular proteins present in selected tissue types under specific conditions, whereas protein microarrays, yeast two-hybrid, and co-purification technologies, alone or in combination, could be used to discover the protein–protein interactions in rice proteome.

(5) Natural Variation of *O. sativa* and its Relatives

We propose two main aims for this objective: (1) sequencing a core set of *O. sativa* strains and its AA-genome relatives; and (2) develop a comprehensive platform for SNP association study to determine the relationship between phenotype and genotype and to identify functional diversity of agriculturally useful genes.

There is a rich collection of rice germplasm resources. The cultivated rice consists of two species—*Oryza sativa* L., referred to as Asian cultivated rice, and *O. glaberrima* Steud., referred to as African cultivated rice. There are also 20 wild species in the genus *Oryza* (Vaughan, 1994). The International Rice Gene Bank holds more than 105 000 types of Asian and African cultivated rice and around 5000 ecotypes of wild relatives (www.irri.org/GRC/GRChome/Home.htm). In addition, many major rice-producing countries have established national germplasm banks. Collectively, these germplasm collections contain genes that can be used to address a broad range of research objectives and agriculture needs. Another development is the construction of 'core collections'. A core collection, by definition, is to represent the largest genetic diversity with the smallest num-

ber of accessions (Brown, 1989). Therefore, these core collections are 'condensed forms' of the germplasms, and would be very useful for discovering agriculturally useful genes.

Advances in sequencing technology have now made it possible for sequencing a large number of lines with affordable costs. IRFGP should seek to obtain whole genome sequences for a large set of the core collections (~1000 accessions), as well as a number of wild rice species known to be rich sources of genes for biotic and abiotic resistances. Those sequencing efforts should produce densely populated SNP markers in the entire rice genome among key accessions, which can be a powerful resource for phenotype association mapping and QTL (Quantitative Trait Locus) cloning.

Many agronomically important genes have been identified and mapped as QTLs during the last two decades using molecular markers. A number of QTLs for yield traits has recently been cloned using the map-based cloning approach, suggesting that QTL cloning has become more efficient. Association mapping analyses of the genome-wide SNP markers with diverse phenotypes, together with carefully constructed mapping populations, will identify agriculturally useful genes conserved in collections of germplasms at unprecedented speed. Sequence information will also allow identification of allelic diversity of defined genes or QTLs. When combined, these approaches would provide a highly efficient strategy for identifying genes of agricultural importance in the species gene pool.

(6) Bioinformatics, Data Management, and Exchange and Sharing of Information

Achieving the research goals specified by a joint effort like the proposed IRFGP will require significant investment in and development of public accessible bioinformatics tools and databases. A significant effort in this area must be expanded in close coordination with the biological aspects of the project. The specific aims here are to create a comprehensive rice annotation database (cRAD) that also provides data-mining platform for high throughput data analysis by individual researchers. Ultimately, the cRAD that we envision will provide a common vocabulary, visualization tools, and information-retrieval mechanisms that permit integration of all knowledge into a seamless whole that can be queried from any perspective.

Database architecture allowing easy integration with other databases will be an essential component of this effort. Divergent types of data (e.g. mutant libraries, full-length cDNAs, genome sequencing, expression arrays, molecular mapping, phenotyping, together with experimental set-up and growth conditions) will need to be integrated and archived. The ability to generate these datasets will easily outpace the ability to rationally maintain, manage, and extract utility from these data. Hence, there is a critical need to invest in novel data-mining approaches and to also bolster support for integrating current databases. Efforts should also be made to integrate the information generated in rice functional genomics research with other major international undertakings in germplasm programs, such as the Generation Challenge Program, and ones

supported by major funding agencies, such as the Gates and Rockefeller Foundations.

(7) Establishment of the Toolkit for High-Throughput Knowledge-Based Rice Breeding

The ultimate goal of the rice functional genomics research is to realize the ideal situation of ‘breeding by design’ to breed cultivars to meet the diverse needs of global rice production for high yield, superior quality, multiple resistances and high nutrient-use efficiency. The full version of ‘breeding by design’ should be composed of four different levels: the yield limit to be achieved through a population structure that can make maximum use of the solar energy in given ecological conditions; the plant architecture to realize the population structure; the traits to make up the plant architecture to achieve high yield, superior quality, resistances to multiple biotic and abiotic stresses, and high nutrient-use efficiency; and the genes to produce the traits.

The knowledge and technologies generated in the course of this research will greatly elevate the level of crop breeding in general, and rice breeding in particular. IRFGP will develop mechanisms to allow dissemination of research findings to breeders and integration of functional genomic research, especially gene identification, with breeding activities.

High throughput and low-cost technologies based on the massive sequence information should be developed for breeding applications, most likely as multiple sets of oligonucleotide chips to meet the diverse needs of rice breeding programs, such as *indica* vs *japonica*, and inbreds vs hybrids, in different countries and regions. Eventually, breeding will become a process of assembling according to designed blueprints.

COMMUNITY DEVELOPMENT

Scientists in the world working on rice biology and biotechnology have evolved into a community during the past thanks to the support of Rockefeller Foundation’s International Program on Rice Biotechnology, and the genome sequencing activities of IRGSP. The changing paradigm of functional genomics will require new types of organization to encourage and facilitate lateral, interdisciplinary approaches to problem solving. Many components of the functional genomics research will be beyond the scope of individual labs; resources and information developed by various institutions need to be shared among the investigators. Success of this project would be critically dependent on the ability to coordinate the activities, to reduce redundancy, and to avoid unnecessary competition.

To facilitate the development and implementation, a number of IRFGP Centers need to be identified and established to take the leadership during the course of the project. The proposed Centers should be self-sustained with reasonable geographical representation. The Centers should play major roles in generating, maintaining and disseminating resources and enabling technologies, and have the ability to solve a wide

range of specific biological problems. Individual investigators will be connected with the Centers in various ways to form network relationships to solve specific problems. The value of this project therefore depends on the ability of the Centers and individual research laboratories to leverage investment from their own governments and other funding agencies.

INTERNATIONAL COOPERATION AND COORDINATION

An International Rice Functional Genomics Steering Committee (IRFGSC), made up of representatives of countries with ongoing major rice functional genomics programs, is already in existence. The primary goals of this committee at present are forging relationships and fostering communication among the involved groups. To fulfill the goals of the RICE2020 Project, the IRFGSC has to undertake stronger responsibilities in leadership and coordination. A number of responsibilities could be undertaken by this committee:

- to coordinate programmatic aspects of the rice functional genomic research worldwide;
- to facilitate open communication and free exchange of data, materials, and ideas in the research community;
- to monitor and summarize progress of scientific activities of participating groups;
- to identify needs and opportunities of the research community and communicate them to funding agencies of participating nations;
- to periodically update and adjust the course of the project.

To fulfill those important responsibilities, IRFGSC will meet face to face at least once a year in conjunction with the International Symposium on Rice Functional Genomics and produce an annual report of the overall progress and status of the research to the community at large.

ACKNOWLEDGMENTS

Development of the ideas for RICE2020 has been in consultation with Drs Rod Wing and Hei Leung.

REFERENCES

- Brown, A.H.D.** (1989). Core collections: a practical approach to genetic resources management. *Genome*. **31**, 818–824.
- Hirochika, H., et al.** (2004). Rice mutant resources for gene discovery. *Plant Mol. Biol.* **54**, 325–334.
- International Rice Genome Sequencing Project** (2005). The map-based sequence of the rice genome. *Nature*. **436**, 793–800.
- Jeon, J.S., et al.** (2000). T-DNA insertional mutagenesis genomics in rice. *Plant J.* **22**, 561–570.
- Liu, X.H., et al.** (2007). A collection of 10,096 *indica* rice full-length cDNAs reveals highly expressed sequence divergence between *Oryza sativa indica* and *japonica* subspecies. *Plant Mol. Biol.* **65**, 403–415.

- Schwab, R., Ossowski, S., Riester, M., Warthmann, N., and Weigel, D. (2006). Highly specific gene silencing by artificial microRNAs in Arabidopsis. *The Plant Cell*. **18**, 1121–1133.
- The Rice Annotation Project (2007). Curated genome annotation of *Oryza sativa* ssp. *japonica* and comparative genome analysis with *Arabidopsis thaliana*. *Genome Res.* **17**, 175–183.
- The Rice Annotation Project Consortium (2008). The Rice Annotation Project Database (RAP-DB): 2008 update. *Nucleic Acids Res.* **36**, D1028–D1033.
- The Rice Full-Length cDNA Consortium (2003). Collection, mapping, and annotation of over 28,000 cDNA clones from *japonica* rice. *Science*. **301**, 376–379.
- Vaughan, D.A. (1994). *The Wilde Relatives of Rice: A Genetic Resource Handbook* (Manila, Philippines: International Rice Research Institute).
- Warthmann, N., Chen, H., Ossowski, S., Weigel, D., and Hervé, P. (2008). Highly specific gene silencing by artificial miRNAs in rice. *PLoS One*. **3**, e1829.
- Wu, C., et al. (2003). Development of enhancer trap lines for functional analysis of the rice genome. *Plant J.* **35**, 418–427.
- Xie, K., Zhang, J., Xiang, Y., Feng, Q., Han, B., Chu, Z., Wang, S., Zhang, Q., and Xiong, L. (2005). Isolation and annotation of 10828 putative full length cDNAs from indica rice. *Sci. China Ser. C Life Sci.* **48**, 445–451.