The Rice Genome: Implications for Breeding Rice and Other Cereals

Yunbi Xu[1] and **Qifa Zhang**

[1]Department of Plant Breeding, Cornell University, Ithaca, NY 14853-1901. www.cornell.edu Email: yx17@cornell.edu
[2]National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, Wuhan 430070, China. Email: qifazh@mail.hzau.edu.cn

Abstract
Rice has been serving as a model crop for cereals in genomics. The availability of complete genome sequences, together with various genomic resources fully developed for both rice and *Arabidopsis*, has revolutionized our understanding of genetic make-up of crop plants. Both macrocolinearity revealed by comparative mapping and microcolinearity revealed by sequence comparisons indicate that sequencing and functional analysis of rice genome will have great impact on other cereals. High-throughput capability, mutant libraries, and advanced transformation technique make functional genomics in rice and other cereals more manageable than ever. Sequences of rice genome and genes have been used to develop functional and biallelic markers that are more useful in genetic mapping and marker assisted selection. It is expected that an integrated database, which is for all rice-related information required for plant breeding and combined with other cereal databases, will be key to the utilization of all genomics resources for plant breeding.

Media summary
Rice genomics and sequences will help understand functions of all the genes in rice and other cereals and thus accelerate breeding programs.

Key Words
Genome sequencing, functional genomics, molecular markers, plant breeding, rice, cereals

Introduction

The genomics revolution of the past decade has greatly improved our understanding of the genetic make up of living organisms including crop plants. Together with the achievements represented by complete genomic sequences of *Arabidopsis* (The Arabidopsis Genome Initiative 2000) and rice (Goff et al. 2002; Yu et al. 2002), high-throughput and parallel approaches are available for the analysis of transcripts, proteome, insertional mutants and chemically induced mutants. All this information allows us to understand the function of genes and the related phenotype. As a large and diverse set of agronomically important crops from the family Gramineae, cereals serve as the main source of dietary calories for most human populations, either consumed directly, as is rice, or indirectly by way of animal feed. The structures of the cereal genomes and the genes contained within them will aid geneticists and molecular biologists in their quest to understand cereal biology and help plant breeders in their goal of developing better products.

It is believed that the cereal species began to diverge 50–70 million years ago. For the past few thousand years, these species have undergone largely parallel selection regimes associated with domestication and improvement. The rice genome sequence provides a platform for organizing information about diverse cereals, and together with genetic maps and sequence samples from other cereals is yielding new insights into both the shared and the independent dimensions of cereal evolution. Sequence capturing techniques promise to accelerate gene discovery in many large-genome cereals, and to better link the under-explored genomes of ‘orphan’ cereals with state-of-the-art knowledge. Genomics-based approaches are used to identify genes that have been involved in cereal improvement. This article discusses the rice genomics and its implications for plant breeding in rice and other cereals.

Rice as a model plant for cereals

In addition to being a major food source, rice is the foundation stone of cereal genomics. Rice is emerging as a model cereal for molecular biological studies because of the following reasons. (1) Rice has a large research community and exceptional agricultural importance, being one of the most important food crops for over half the world’s population; (2) Rice has the smallest genome size (estimated at about 430Mb) among cereal crops, which makes it most manageable at the whole genome level; (3) The highly

26 Sep – 1 Oct 2004, Brisbane, Australia. Published on CDROM. Web site www.cropscience.org.au

conserved gene order and gene content within the cereals indicate that rice research can greatly benefit other grass research programs; (4) The landmark draft sequences of the indica and japonica rice genomes published in 2002 and the more complete genome sequence that will be finished by the end of 2004 provide a powerful new resource for rice studies; (5) Rice is relatively easy for transformation compared to the other major cereals which makes routine use of transgenics possible for a variety of research purposes; (6) Large numbers of genetic stocks have been accumulating across the rice community, including permanent mapping populations, introgression/substitution lines, various genetic mutants, near-isogenic lines, molecular markers, and etc. (7) Several genomic databases have been developed which contribute not only to depository, searching, querying, analyzing of all types of information that have been creating in rice community but also to the comparison with other databases and organisms; (8) International efforts in functional genomics have generated enormous amounts of genomic resources and toolkits, including large T-DNA insertion libraries tagged by flanking sequences, expressed sequence tag (EST) databases, and whole genome cDNA and oligo chips, together with expression profiling techniques, and large collections of full-length cDNA clones, which will greatly accelerate the processes of gene identification.

Whole-genome analysis of rice suggests that the *Arabidopsis* genes appear to be a subset of the genes found in rice. *Arabidopsis* genes, along with genes from other eukaryotes, are useful in assigning functions to cereal genes. However, the extensive rearrangement of rice and *Arabidopsis* genomes indicated that monocot-dicot chromosomal synteny is limited and thus the utilization of *Arabidopsis* genome information for understanding cereal genomes as a whole might also be limited. Because the genomes of cereal species have similar gene content and structure (Goff et al. 2002), the structural and functional genomic analysis of rice will enable similarly wide-scale gene discovery in other cereals. The relatively close relationship among the major cereals suggests that their study and improvement can benefit considerably from the sequences of small-genome relatives such as rice. Whole genome sequencing, together with extensive genetic and physical mapping efforts, provides a foundation for organizing information about diverse cereals, identifying orthologous genes, facilitating the genome sequencing in other cereals, and yielding new insights into both the shared and the independent dimensions of cereal evolutionary history. A finished rice genome will provide a complete index of potential rice genes but will not tell us which of these genes are important in providing desired traits in cereal crops. In the future, techniques such as gene, protein and metabolic profiling will provide insights into the function and expression patterns of genes and into how these genes ultimately contribute to a crop's ability to react to an environment and reproduce (Ware and Stein 2003). In addition, this treasure of sequence information can be used to acquire an almost unlimited number of DNA markers and genes for crop improvement. The full set of genes not only permits comprehensive characterization of gene expression by several high-throughput approaches but also inspection of the gene complements in rice and related species to see which pathways are shared and which are unique, and how these pathways may have been modified. Rice sequence information will help gain instantaneous access to and monitor the genes in breeding populations and help evaluate unlimited cereal germplasm through novel types of molecular markers such as single nucleotide polymorphisms (SNPs) and intragenic or functional markers. It also provides unlimited opportunities to relate specific changes in gene structure and content to identify the differences in different plant, animal and microbial species. Using rice database for complete set of predicted and known peptides, those that are of most interest for three-dimensional characterization can be identified. Finally, rice sequence information can be used to mine rice genome and genes to understand how gene families are created, amplified and diverge to create new biological activities and specificities.

In addition to the whole genome sequence, the availability of more than 300,000 public ESTs and a large set of full-length cDNA clones make the annotation of the rice genome a slightly less daunting task. The information on the full-length cDNAs will be extremely important for functional genomics and proteomics of rice, and because no complete full-length cDNA data are available for any other cereal genomes, it will have a great impact on future studies of plant genomics in general (Shimamoto and Kyozuka 2002). Full-length cDNA clones are also necessary to identify exon-intron boundaries and gene-coding regions within genomic sequences and for comprehensive gene-function analyses at the transcriptional (transcriptomic) and translational (protein informatic) levels (The Rice Full-Length cDNA Consortium 2003). The availability of ESTs from a diverse set of cDNA libraries provides information on the transcript abundance, tissue location, and developmental expression of genes.

Colinearity among rice and other cereals

Macrocolinearity

Significant genomic colinearity in plants has been revealed by comparative genetic mapping and genome sequencing, although plant genomes vary tremendously in genome size, chromosome number, and chromosome morphology. Comparative mapping of cereal genomes using low copy number, cross-hybridizing genetic markers has provided compelling evidence for a high level of conservation of gene order across regions spanning many megabases (i.e. macrocolinearity). Initial studies of the organization of grass genomes indicated that individual rice chromosomes were highly colinear with those of several other grass species, and extensive work over the past decade has shown a remarkably consistent conservation of large segments of linkage groups within rice, maize, sorghum, barley, wheat, rye, sugarcane, and other agriculturally important grasses (e.g. Ahn and Tanksley 1993; Kurata et al. 1994; Wilson et al. 1999). These studies led to the prediction that grasses could be studied as a single syntenic genome. Therefore, if the ortholog of a studied gene could be confirmed by comparative genetic maps, then knowledge acquired from one species could be compared to the results of similar experiments in another species. The macrocolinearity was summarized by Gale and Devos (1998) for rice, oats, maize, sorghum, sugar cane, foxtail millet, wheat, and finger millet using what is now known as the 'Circle Diagram'. The initial work on the colinearity of genetic markers was reinforced when QTL controlling important agronomic traits were also mapped to collinear regions among grass genomes (Paterson et al. 1995; Peng et al. 1999; Chen et al. 2003).

This unified grass genome model has had a substantial effect upon plant biology, but has not yet lived up to its potential. There appear to be two major reasons for the relatively slow application of this approach. First, genomic sequence data are largely lacking for grass species other than rice. Second, the colinearity of gene order and content observed at the recombinational map level is often not observed at the level of local genome structure (Bennetzen and Ramakrishna 2002; Feuillet and Keller 2002). The frequency of major chromosomal rearrangements observed among genomes depends on the degree of relatedness of the species investigated.

Microcolinearity

Despite the general colinearity exhibited by comparative genetic maps, rearrangements that involve regions smaller than a few cM may occur and would be missed by most recombinational mapping studies. Comparative sequence analysis involving large genomic segments can detect these rearrangements. Such analyses reveal the composition, organization, and functional components of genomes and provide insight into regional differences in composition between related species. Recently, the sequencing of long regions of the cereal genomes has allowed microcolinearity across gene clusters to be investigated. Several recent studies in the cereals have demonstrated incomplete microcolinearity at the sequence level (Tarchini et al 2000; Dubcovsky et al. 2001; Song et al. 2002; SanMiguel et al. 2002; Tikhonov et al. 2000). Song et al. (2002) identified orthologous regions from maize, sorghum, and two subspecies of rice. It was found that gross macrocolinearity is maintained, but microcolinearity is incomplete among these cereals. Deviations from gene colinearity are attributable to micro-rearrangement or small-scale genomic changes, such as gene insertions, deletions, duplications, or inversions. In the region under study, the orthologous region was found to contain six genes in rice, 15 genes in sorghum (of which three have been amplified, producing a total of 29 genes), and 13 genes in maize (of which one has been amplified, resulting in a total of 34 genes). In maize and sorghum, gene amplification caused a local expansion of conserved genes but did not disrupt their order or orientation. As indicated by Bennetzen and Ma (2003), numerous local rearrangements differentiate the structures of different cereal genomes. On average, any comparison of a ten-gene segment between rice and a distant grass relative such as barley, maize, sorghum or wheat shows one or two rearrangements that involve genes. A simple extrapolation to the rice genome of about 40 000 genes (Goff et al. 2002) suggests that about 6000 genic rearrangements will differentiate rice from these other cereals. Most of these rearrangements appear to be tiny and thus would not interfere with the macrocolinearity observed by recombinational mapping. There are exceptions, however, which include chromosomal arm translocations and movements of single genes to different chromosomes.

As expected, there is a high degree of gene conservation between the two shotgun-sequenced subspecies of rice, japonica and indica, which diverged more than 1 million years ago. On careful inspection, however, narrow regions of divergence can be found in these genomes (Song et al. 2002). These regions correspond to areas of increased divergence among rice, sorghum and maize, suggesting that the

alignment of the two rice subspecies might be useful for identifying regions of cereal genomes that are prone to rapid evolution. Similar comparative analyses of *Arabidopsis* accessions have shown that both the relocation of genes and sequence polymorphisms between accessions (in both coding and non-coding regions) are common in the *Arabidopsis* genome (The *Arabidopsis* Genome Initiative 2000; Rossberg et al. 2001). Intraspecific violation of genetic colinearity has also been identified in maize (Fu and Dooner 2002). Han and Xue (2003) showed extensive conservation of microcolinearity in gene order and gene content between indica and japonica, but they also discovered significant numbers of rearrangements and polymorphisms when comparing the two genomes. The deviations from colinearity are frequently owing to insertions or deletions. Intraspecific sequence polymorphisms commonly occur in both coding and non-coding regions. These variations often affect gene structures and may contribute to intraspecific phenotypic adaptations.

One of the standard and most powerful tools of molecular biology is the ability to efficiently compare the sequence of any gene with the sequences of all previously characterized genes. Comparative genetics has been facilitated by the development of massive databases, efficient query and comparison software, and ever-improving computers. Many of the first genes to be sequenced in rice and other grasses were represented by abundant mRNAs (e.g. those encoding storage proteins and photosynthetic proteins). Members of the same gene families (e.g. paralogs), including those that were mapped to the same genomic position and thus were derived by vertical descent from a common ancestral gene (i.e. orthologs), were often cloned and analyzed in multiple species (Bennetzen and Ma 2003). Comparisons of gene family members within and between species yielded the expected result that the genes were most highly conserved between the most closely related species. Moreover, sequence conservation was greatest in the protein-coding portions of the exons.

Implications of genome colinearity

Genomics would be much simpler if the order of genes were common (syntenic) across major groups of plants. As the analysis of the *Arabidopsis* sequence provides information that will facilitate the annotation of the rice sequence and *Medicago* provides a resource for research on some legumes, the effort put into sequencing and annotating rice genomes will be well rewarded: annotation will be transferred to related sequences and used again and again. Although other cereal genomes will not be sequenced soon, the synteny between the monocots will help decipher the structure and function of the more complex genomes. A fully assembled rice sequence allows the more accurate assessment of rice's macro- and microsynteny with other cereals. Although these studies are still in their early stages, many exceptions and breakages in synteny have been observed, as discussed above, when comparing various cereal genomes (Tarchini et al. 2000; Song et al. 2002; Bennetzen and Ma 2003). Rice genome information will provide new innovations in rice research, as well as a huge amount of new knowledge, tools and opportunities for plant genome biology in other cereals.

Map-based cloning in plants of large genome sizes, such as barley, maize, and wheat, has been extremely difficult. However, it may be easier to use comparative maps to isolate a mapped gene from a large genome using a related plant with a small genome. Markers linked to the gene of interest and prior knowledge about colinearity of this region between large and small genomes are essential to isolate a gene using this approach. Under almost all circumstances, a small genome species will provide numerous DNA markers on a single BAC, which permits more detailed mapping in the large genome species. Chromosome walking involves identifying low copy number DNA markers that are tightly linked to the gene of interest and using them as probes to screen large insert BAC libraries to identify appropriate clones. Repeated rounds of such screening using low copy number regions from a series of BACs may be required to identify overlapping clones extending toward the target gene. However, this approach has potential pitfalls, especially with respect to some disease resistance genes, because resistance gene regions often undergo rapid rearrangement that results in a lack of microlinearity caused by deletion or translocation of the target loci. Probably the most comprehensive application of colinearity between rice and another cereal species was the attempt to clone specific barley disease resistance genes by chromosome walking in rice. The colinearity provided numerous DNA markers from rice that facilitated the chromosome walk in barley, leading to the isolation of the desired resistance genes (Brueggeman et al. 2002).

The cross-utilization of information from botanical models such as rice, in the study and improvement of major cereal crops requires a detailed understanding of the evolutionary history of cereal genomes. Toward a complete picture, broadening knowledge of diversity and extending data from botanical models across (and beyond) the cereals are required and domestication and improvement need to be footprinted. One important outcome of comparative gene analysis among rice and other cereals is the insight that will be gained into the evolutionary aspects of gene functions in higher plants. Whole-genome duplication, through polyploidization, segmental duplication, and local gene amplification, increases the number of paralogous gene sequences found in plants. Together, these forms of duplication explain why the numbers of genes in grass genomes exceed those in the other eukaryotic genomes sequenced so far, including the human genome. The use of all three duplication mechanisms in the evolution of grass genomes suggests that grasses, like other plants, may have evolved more rapidly than could be predicted by comparing only coding-sequence substitution rates. This rapid evolution may be one of the reasons for the degree of speciation within certain large families of Gramineae. It will be interesting to determine in evolutionary terms how the cereal genomes have maintained a level of macro-colinearity in light of the dynamic plant genome.

Where microcolinearity is broken and a gene that is present in one cereal is 'missing' from its orthologous position in another, it is often possible to find a matching gene homologue in a non-orthologous location (Song et al. 2002; Xu et al. 2002). The putative mechanism for this phenomenon is an ancient gene duplication in the common ancestor followed by the loss of one gene copy in the first modern species and the loss of the other copy in the second species. Detailed analyses of the genomes of several model organisms revealed that large-scale gene or even entire genome duplications have played a prominent role in the evolutionary history of many eukaryotes. Recently, strong evidence has been presented that the genomic structure of the dicotyledonous model plant species *Arabidopsis thaliana* is the result of multiple rounds of entire genome duplications. By analyzing the genome of the monocotyledonous model plant species rice (*Oryza sativa*), a substantial fraction of all rice genes (approximately 15%) were found in duplicated segments (Vandepoeple and Simillion 2003). Dating of these block duplications, their nonuniform distribution over the different rice chromosomes, and comparison with the duplication history of *Arabidopsis* suggest that rice is not an ancient polyploid, as suggested previously, but an ancient aneuploid that has experienced the duplication of one-or a large part of one-chromosome in its evolutionary past, approximately 70 million years ago. This date predated the divergence of most of the cereals, and relative dating by phylogenetic analysis showed that this duplication event is shared by most if not all of them.

From sequence to gene function

As we trek into the uncharted territories of the genomic era, there is an urgency for the development of approaches for assigning functions to the multitude of uncharacterized genes. Arabidopsis genomics and functional genomics have led the way in this regard and will provide the blueprint for the structure and function of many plant genes. Some of the remarkable insights emerging from whole-genome analysis are the number of gene clusters, large-scale duplications of chromosome segments, as well as surprisingly high frequency (40%) of newly discovered genes of unknown function. In this section, several important issues related to functional analysis of genes will be discussed.

High-throughput techniques

One of the challenges in functional genomics is to understand how thousands of gene products interact with each other to control development and the ability of an organism to respond to its environment. Challenge to determine the function of the tens of thousands of plant genes, many of them showing no detectable homology to genes for which cellular roles have been identified in bacteria, yeast, or animals, has triggered an advantage of novel methodologies. DNA-based microarrays that detect the accumulation of transcripts from thousands of genes in a single hybridization experiment are one of the tools available to help meet this challenge. High-throughput techniques such as oligo chips, gene chips and various serial analysis of gene expression (SAGE) techniques are used for global gene expression analysis at a particular stage or time. The gridding of thousands of unique DNA sequences in large or small arrays provides a substrate that can be used to identify candidate genes that exert an influence at specific points in development. Common sources of DNA for the arrays include cDNA, ESTs, subgenomic regions of specific chromosomes, and even the entire set of genes in *Arabidopsis*. A microarray technology for rice gene-expression studies has been developed and is now applied in a number of studies. Oligonucleotide

based probe array (Gene Chip) technology, which has been commercially applied for *Arabidopsis* genes, has also been developed in rice (Q. Zhang, unpublished). Availability of these technologies for various gene expression studies will be essential for future studies on the functions of rice genes.

Map-based cloning

Although map-based cloning is feasible and many important rice genes have been isolated for disease resistance, heading date, and semidwarfism, it is too time-consuming and laborious for the molecular identification of genes, especially quantitative trait loci (QTL). Map-based gene cloning has been reshaped because of the availability of sequence information and new gene mapping techniques such as linkage disequilibrium/association mapping. The chromosome-aligned genome sequence information allows skip several of the steps in map-based cloning (Jander et al. 2002). A certain level of genetic mapping may be able to associate the target trait to a specific genomic region with a lager number of sequence-based molecular markers available, and further fine mapping effort may narrow the target genomic region to several gene candidates based on sequence information. This was followed by cloning, complementation by transformation, and de novo determination of the sequence of the entire region of interest to high quality without a previously determined wild-type DNA sequence as a guide.

Functional genomic approaches

Functional genomics has been broadly applied to include many endeavors aimed at determining functions of genes on a genome-wide scale, such as sequence alignment-based comparisons to identify homologs between and within organisms; transcriptional profiling to determine gene expression patterns; and yeast two-hybrid and other interaction analyses to help identify pathways, networks, and protein complexes (Henikoff and Comai 2003). The *Arabidopsis* community has developed an initiative to empirically identify the function of every *Arabidopsis* gene by the year 2010. Although a daunting task, several approaches have already been established, including the use of T-DNA knock-out lines, overexpression studies, and denaturing high power liquid chromatography. In contrast to the previously prevalent gene-by-gene approaches, new high-throughput methods are being developed for expression analysis as well as for the recovery and identification of mutants. The experimental approach is consequently changing from hypothesis-driven to nonbiased data collection and an archiving methodology that makes these data available for analysis by bioinformatics tools. The functional genomics methodology is also changing the experimental strategy from a forward genetics (mutant to gene) approach to a reverse genetics (sequenced gene to mutant and function) approach.

In functional genomics, libraries or populations of mutants that cover all possible genes becomes an increasingly important tool. Mutant libraries are being constructed in rice and other cereals using chemical and physical mutagenesis, T-DNA insertion, and transposon tagging, which can be used for functional analysis based on the loss-of-function analyses (Hirochika et al. 2004). Gain-of-function approaches such as T-DNA activation tagging and gene overexpression are powerful complements to insertional mutagenesis that creates formidable obstacles to the successful identification of mutant phenotype. With microarray techniques widely used in cDNA-based expression profiling, transcript profiling, and metabolite profiling, phenotypic profiling is playing important role in functional analysis of plant genes. Although currently available knock-out methodologies could be used for uncovering the function of newly discovered genes, the mixed outcomes in terms of the success of these approaches in down-regulating gene expression necessitate the development of new functional genomics tools.

The identification of genes by computational approaches is relatively straightforward for organisms with compact genomes (such as bacteria and yeast), because exons tend to be large, and the introns are either nonexistent or short. The challenge is much greater for larger genomes (such as those of rice and maize), because the exonic “signal” is buried under nongenic “noise.” Computational sequence analysis methods, which detect genes in genomics DNA, can be broadly classified into two main categories: homology-based methods, and *ab initio* methods. Currently, computational methods are usually exploited complementary to other functional genomics approaches.

It is expected that the functional genomics of model plants will contribute to the understanding of basic plant biology as well as the exploitation of genomic information for crop improvement. This is because a large number of gene functions for generic traits will be functional across species, either directly or after identifying the functional homologues. Perhaps the most exciting application of comparative cereal

genomics will be the identification of different versions of rice genes from other species. Orthologous genes in other cereals will be similar in sequence and function to those in rice but could result in markedly different phenotype. These genes will be available for introduction into rice to produce new types of plants with many novel features.

Transformation

Transformation of allelic series into isogenic genetic backgrounds can confirm the function of individual sequence motifs. However, current plant transformation protocols based on nonhomologous end joining result in random genomic integration of transgenic DNA, position effects, multiple insertions of the transgene and transgene alterations (Hanin and Paszkowski 2003), obscuring quantitative phenotypic differences between alleles. This can be circumvented using homologous recombination (HR)-based, locus-targeted integration of alleles. Recently, 1% of insertion events in rice (*Oryza sativa*) were found to result from HR (Terada et al. 2002). If this proves correct, rice genomics-genetics will be revolutionized. Further, if the method can be applied to other species, a similar advance in genomics of all plants would occur.

From sequences to molecular markers and marker-assisted selection

Functional markers

Genetic diversity at or below the species level is mostly characterized by molecular markers that more or less randomly sample genetic variation in the genome. This type of "neutral" or random marker (RM) is a very effective tool, amongst others, for the establishment of the breeding system, the study of gene flow among natural populations, and the determination of the genetic structure of genebank collections. RM systems are still the methods of choice for marker-assisted breeding. However, 'users' of biodiversity are often not interested in random variation but rather in variation that might affect the evolutionary potential of a species or the performance of an individual genotype. Such 'functional' variation can be tagged with neutral molecular markers using QTL and linkage disequilibrium mapping approaches. Alternatively, DNA-profiling techniques may be used that specifically target genetic variation in functional parts of the genome.

Different approaches (including association studies) have recently been adopted for the functional characterization of allelic variation in plants and to identify sequence motifs affecting phenotypic variation. Andersen and Lübberstedt (2003) proposed the term 'functional markers' (FMs) for DNA markers derived from such functionally characterized sequence motifs. Functional marker development requires allele sequences of functionally characterized genes from which polymorphic, functional motifs affecting plant phenotype can be identified. Functional markers are superior to RMs such as RFLPs, SSRs and AFLPs owing to complete linkage with trait locus alleles and functional motifs. In contrast to RMs, FMs allow reliable application of markers in populations without prior mapping, the use of markers in mapped populations without risk of information loss owing to recombination, and better representation of genetic variation in natural or breeding populations. Once genetic effects have been assigned to functional sequence motifs, FMs derived from such motifs can be used to fix gene alleles (defined by one or several FM alleles) in several genetic backgrounds without additional calibration. This would be a major advance in marker applications, particularly in plant breeding, to select (for example) parent materials to build segregating populations, as well as subsequent selection of inbred lines (Andersen and Lübberstedt 2003). Depending on the mode of FM characterization, FMs can also be used for targeted combination of alleles in hybrid and synthetic breeding and variety testing based on the presence or absence of specific alleles at morphological trait loci. In population breeding and recurrent selection programs, FMs can be used to avoid genetic drift at characterized loci.

Conserved ortholog set (COS) markers

As an challenge to finding the manner in which map, sequence, and eventually functional genomic information from one species can be accessed, compared, and exploited across all plant species, it will require the identification of a subset of plant genes that have remained relatively stable in both sequence and copy number since the radiation of flowering plants from their last common ancestors. Identification of such a set of genes also would facilitate taxonomic and phylogenic studies in higher plants that are based at present on a very small set of highly conserved sequences, especially those of chloroplast and mitochondrial genes. Fulton et al. (2002) screened a large tomato EST database against the *Arabidopsis* genomic sequence and reported the identification of a set of 1025 genes (referred to as a conserved

ortholog set, or COS markers) that are single or low copy in both genomes (as determined by computational screens and DNA gel blot hybridization) and that have remained relatively stable in sequence since the early radiation of dicotyledonous plants. This set of COS markers, identified computationally and experimentally, may further studies on comparative genomes and phylogenetics and elucidate the nature of genes conserved throughout plant evolution.

Single nucleotide polymorphism (SNP)
Comparative information about the chromosome organization of the two closely related rice subspecies has important implications for the development of new molecular markers. Similar comparative analyses of *Arabidopsis* accessions have shown that both the relocation of genes and sequence polymorphisms between accessions (in both coding and non-coding regions) are common in the *Arabidopsis* genome (The *Arabidopsis* Genome Initiative 2000; Rossberg et al. 2001). The forward genetics approach for identifying functionally important genes derives from a known allelic difference conferring an improved phenotype. In such an approach, the objective is to identify a sequence change conferring the improved phenotype. Such a sequence change can then become the basis for a marker that is specific for that allele. These types of markers will always cosegregate with the trait of interest and should also be polymorphic in any cross. Such a marker will often be based on SNP. By systematically searching nucleotide differences, a complete set of markers that is based on SNPs or other sequence variations could be developed. Comparative genetic analysis of rice germplasm can be used to identify the existing genetic variation or multiple alleles by means of SNPs analysis for agronomically important traits using gel based or DNA-chip based methods. In the reverse approach, SNPs are sought in candidate genes to identify the phenotypic effects of genes. Known SNPs can be used to identify new candidate genes through association mapping. Phenotypic differences that correspond to particular SNPs may be the result of the sequence change. These SNPs can then be used for MAS or screening germplasm and elite breeding lines. The nucleotide change that contributes to quantitative variation has been referred to as a quantitative trait nucleotide or QTN. Fine-mapping combined with sequence analysis could narrow the chromosomal region associated with quantitative variation (QTL) down to a specific nucleotide change (Xu 2002).

Germplasm evaluation
The evaluation of germplasm resources is required for the continuous improvement of crop plants. Vast genetic resources are available for rice and other crops, but, to date, few of them have been characterized at the molecular level. Automated, high-throughput genotyping systems make large-scale marker-assisted germplasm evaluation possible. Molecular markers can be used in germplasm evaluation for (1) differentiating cultivars and constructing heterotic groups; (2) identifying germplasm redundancy, underrepresented alleles, and genetic gaps in germplasm collections; (3) monitoring genetic shifts that occur during germplasm storage, regeneration, domestication, and breeding; (4) screening germplasm for novel/superior genes (alleles); and (5) constructing a representative subset or core collection (Xu et al. 2003).

Ideally a breeder would like to know all alleles of all genes across germplasm accessions. Once all alleles are known, a breeder would like to rank them. This will be a complex exercise since an allele of a gene for an agronomic trait would usually not be good or bad in itself, but in the context of other alleles in a genomic network. Next, one could deduce why one allele is better than another. From there one could design synthetic alleles that are better than the ones provided by nature. For the coding region of genes this could be done randomly (e.g. by gene shuffling) or via a targeted approach (e.g. by domain swapping). So genomics is going to bring to the breeder a better description (including allele-tagging) as well as an extension of his breeding material. The bottleneck of all of this is reliably assessing the phenotypes caused by countless different alleles and allele combinations using available germplasm resources. Molecular methods, on the other hand, can be used to screen a large number of accessions through a pooling strategy. This can be used to screen germplasm collections for alleles of candidate genes that are involved in important processes of the plant, even though known variants for these genes have not been observed through genetic studies. A DNA bank is currently being developed for a core collection of the rice gene bank at IRRI to undertake allele mining.

Marker-assisted selection
Marker-assisted selection (MAS) can be considered the first benefit that breeders can obtain now from genomics. With existing techniques, however, the use of molecular markers is still quite expensive for

application on a large scale in rice breeding programs. Use of MAS will be more beneficial for specific applications. There are six situations suitable for MAS with the current knowledge available (Xu 2002). These include selection without testcrossing or a progeny test; selection independent of environments; selection without laborious fieldwork or intensive laboratory work; selection at an earlier breeding stage; selection for multiple genes and/or multiple traits; and whole genome selection. Examples of where MAS would be advantageous include selection for traits that are difficult or expensive to measure (e.g., salt tolerance, restorer genes); pyramiding multiple genes that confer a similar or identical phenotype (e.g., multiple genes for resistance to blast or bacterial blight in rice); or selecting against the donor chromosomal segments in a backcrossing scheme. Using rice and other cereal crops as examples, Xu (2003) provided a comprehensive review on MAS system, germplasm evaluation, hybrid prediction, and seed quality control.

Novel alleles and genetic diversity widely exist in wild relatives of cultivated plants. For example, wild relatives of rice within the genus *Oryza* are not only a rich source of information on the origins of variation within the genus but also a viable source of a wide variety of agronomically important germplasm for future breeding. To fill the gulf between national research programs and breeding applications in developing countries, an international program, Challenge Programs (www.cerealgenomics.org), was established to unlock the genetic diversity existing in a wide spectrum of germplasm collections. Molecular markers have been proven particularly useful for accelerating the backcrossing of a gene or QTL from exotic cultivars or wild relatives into an elite cultivar or breeding line. Favorable genes or alleles from wild species of rice have been detected after backcrossing to elite cultivars (Xiao et al. 1998; Moncada et al. 2001). Similarly, this approach can identify alleles from exotic cultivars that result in improved phenotype, even though the parent may not possess inferior phenotype for this trait. This approach is thought to be promising in rice because a number of rice cultivars are widely grown for their adaptation, stable performance, and desirable grain quality. Chen et al. (2000) used such an approach to transfer the bacterial blight resistance gene *Xa21* into Minghui 63, a widely used parent for hybrid rice production in China. Ahmadi et al. (2001) used a similar approach to introgress two QTLs controlling resistance to rice yellow mottle virus into the cultivar IR64. Such approaches, however, can only sample a small number of accessions.

Sequence and bioinformatics

As rice genome data will increase exponentially in the areas of genetic mapping, genome sequencing, gene expression monitoring, insertional mutagenesis, and map-based cloning, adequate tools for input, integration, and query will become necessary. Eventually integrating information on rice structural and functional genomics will provide an overall view of the network of genes involved in complex biological responses. Rice genome databases that evolve from rigorous and systematic sequencing efforts should not merely function as storehouses for thousands of bases or amino acids. Of particular importance is the ability to attach substantial genomic information to the sequence. These databases should therefore provide the framework to allow postsequencing analysis such as identifying genes and predicting the proteins they encode, determining when and where the gene proteins are expressed and how they interact, and how these expression and interaction profiles are modified in response to environmental signals. One way to address this need is to interlink the resources of various types of information such as genomic data, phenotypic or expression data, and genetic resources. For a given gene, the database would horizontally link sequence, structure, and map position and would connect related elements of the same type pertaining to the expression profile, protein, and phenotype. All this information should be defined in terms of the genetic resources available for rice and other cereals. Logical connections to other information will enhance the intrinsic value of the genomic data to facilitate new biological discoveries and simulate approaches for effective cereal improvement. Comparative bioinformatics will offer possibilities to link various cereal crops through their genomes and provide keys to understanding how genes and genomes are structured, how they function, and how they evolved. As genetic mapping as well as some preliminary sequence data show the extent of synteny among cereal crops, the creation of links between different databases may foster interoperability, and linkage and interactions should also be promoted between databases of rice and other organisms.

Rice, being a model system for other grass crops, should establish an informatics infrastructure designed to interlink database resources on rice genomics to better serve a more focused research community using this system in contrast to a lager user community. The interlinking of database resources should be

extended to databases of other cereal crops as well. One of the most serious challenges of specialized or expert domain databases, best presented by model organism databases, is to balance the needs between the broader scientific community and the specialized focused groups. To provide the cereal community with a resource for applying the rice genome information to other species, a system was established for making curated and sequence-based correspondences between genetic and physical maps (www.gramene.org). This resource includes curated correspondences between the MMP BAC based maize finger print contig maps based on genetic markers, expressed sequence tagged clusters, and sequence tagged BAC clones anchored to a maize BAC assigned to a finger print contig. The correspondences between the maize and rice genome are available as a graphical display as well as a downloadable tabular format.

Perspectives

There is a gulf between genomics and its application to plant breeding although there are many opportunities available. Plant breeders have to work with a large number of agronomic traits (most quantitatively) under a wide range of environments during a relatively short period. As a result, genomics can be applied to plant breeding only when an integrated package becomes available that combines multiple components such as high-throughout techniques, cost-effective protocols, global integration of genetic and environmental factors, and precise determination of quantitative trait expression.

Functional genomics has become more practical because of advances in science and technology. Some examples are DNA-capturing techniques to facilitate gene hunting efforts in highly-repetitive genomes, motif-directed profiling to specifically target genetic variation in functional parts of the genome, gene expression profiling to identify Expression Quantitative Trait Loci (eQTLs), and RNAi to silence individual genes without altering the genome structure. Microarrays or other nongel systems may allow whole-genome analysis of large number of plants commonly grown in breeding programs. For example, proteome chips can be used to perform various biochemical analyses, protein-protein interaction assays, protein-DNA/RNA interactions, protein-phospholipids interactions and the identification of substrates for kinases and other enzymes. On the other hand, the global analysis at levels of genome (discovery of all genes), transcriptome (quantification of genes expression), proteome (cataloguing of all proteins) and metabolome (estimation of all types of metabolites) has become possible. To assign cellular function of many novel genes which are predicted from the whole genome analysis, several high-throughput approaches, such as DNA chips, SAGE, RNA-mediated interference, gene traps, yeast two-hybrid screening and metabolites quantification, can be employed in functional genomics of rice and other cereals. Ultimately, the goal will be to rapidly assay the genetic makeup of individual plants or varieties in breeding populations.

As a general strategy for identifying induced point mutations, targeting induced local lesions in genomes (TILLING) should be briefly mentioned here. TILLING is a reverse genetic method that combines random chemical mutagenesis with PCR-based screening of gene regions of interest (McCallum et al. 2000). This provides a range of allele types, including missense and knock-out mutations, which are potentially useful in a variety of gene function and interaction studies. Direct proof of sequence motif function can be obtained by comparing isogenic genotypes differing in single sequence motifs. Combined with phenotyping, TILLING provides direct proof of function of both induced and natural polymorphisms.

Many novel techniques need to be improved before they can be widely used in functional genomics analysis and plant breeding. For example, insertional mutagenesis has been recognized as the potential to provide a vast catalog of knockout mutations for an organism. However, a significant insertion-site bias, plus a high level of background mutation, can confound subsequent phenotypic analysis. Also the efficacy of mutagenesis screens in identifying gene functions is limited because of the majority of genes displaying no obvious phenotype and other methodological considerations. Insertions might not result in null alleles, depending on their position within the open reading frame, or intronic or untranslated regions, and effort is required to confirm that any resulting phenotype is not because of unlinked independent insertions. Because of genetic redundancy, a vast proportion of genes are silent when knocked out, or might only show a subtle phenotype or one only revealed in extreme environmental conditions. With all advances in functional genomics and genome sequencing, it is expected that plants should become more manipulable at both phenotypic and molecular levels with more directed breeding objectives. However,

new tools and technologies developed in genomics will greatly enhance, but not replace, the conventional breeding process.

References

Ahmadi N, Albar L, Pressoir G, Pinel A, Fargette D and Ghesquiere A (2001) Genetic basis and mapping of the resistance to rice yellow mottle virus. III. Analysis of QTL efficiency in introgressed progenies confirmed the hypothesis of complementary epistasis between two resistance QTLs. Theoretical and Applied Genetics 103,1084-1092.

Ahn S and Tanksley SD (1993) Comparative linkage maps of the rice and maize genomes. Proceedings of the National Academy of Sciences of the United States of America 90,7980-7984.

Andersen JR and Lübberstedt T (2003) Functional markers in plants. Trends in Plant Science 8,554-560.

Bennetzen JL and Ramakrishna W (2002) Numerous small rearrangements of gene content, order, and orientation differentiate grass genomes. Plant Molecular Biology 48,821-827.

Bennetzen JL and Ma J (2003) The genetic colinearity of rice and other cereals on the basis of genomic sequence analysis. Current Opinion in Plant Biology 6,128-133.

Brueggeman R, Rostoks N, Kudrna D, Kilian A, Han F, Chen J, Druka A, Steffenson B and Kleinhofs A (2002) The barley stem rust-resistance gene *Rpg1* is a novel disease-resistance gene with homology to receptor kinases. Proceedings of the National Academy of Sciences of the United States of America 99,9328-9333.

Chen H, Wang S, Xing Y, Xu C, Hayes PM and Zhang Q (2003) Comparative analyses of genomic locations and race specificities of loci for quantitative resistance to *Pyricularia grisea* in rice and barley. Proceedings of the National Academy of Sciences of the United States of America 100,2544-2549.

Chen S, Lin XH, Xu CG, Zhang QF (2000) Improvement of bacterial blight resistance of 'Minghui 63', an elite restorer line of hybrid rice, by molecular marker-assisted selection. Crop Science 40,239-244.

Dubcovsky J, Ramakrishna W, SanMiguel P et al (2001) Comparative sequence analysis of collinear barley and rice BACs. Plant Physiology 125,1342-1353.

Feuillet C and Keller B (1999) High gene density is conserved at syntenic loci of small and large grass genomes. Proceedings of the National Academy of Sciences of the United States of America 96,8625-8270.

Fu H and Dooner HK (2002) Intraspecific violation of genetic colinearity and its implications in maize. Proceedings of the National Academy of Sciences of the United States of America 99,9573-9578.

Fulton TM, Van der Hoeven R, Eannetta NT and Tanksley SD (2002) Identification, analysis, and utilization of conserved ortholog set markers for comparative genomics in higher plants. Plant Cell 14,1457-1467.

Gale MD and Devos KM (1998) Comparative genetics in the grasses. Proceedings of the National Academy of Sciences of the United States of America 95,1971-1974.

Goff SA, Ricke D, Lan TH et al. (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). Science 296:92-100.

Han B and Xue Y (2003) Genome-wide intraspecific DNA-sequence variations in rice. Current Opinion in Plant Biology 6,134-138.

Hanin M and Paszkowski J (2003) Plant genome modification by homologous recombination. Current Opinion in Plant Biology 6,157-162.

Henikoff S and Comai L (2003) Single-nucleotide mutations for plant functional genomics. Annual Reviews in Plant Biology 54,375-401.

Hirochika H, Guiderdoni E, An G, Hsing YI, Eun MY, Han CD, Upadhyaya N, Ramachandran S, Zhang Q, Pereira A, Sundaresan V and Leung H (2004) Rice mutant resources for gene discovery. Plant Physiology (in press)

Jander G, Norris SR, Rounsley SD, Bush DF, Levin IM and Last RL (2002) Arabidopsis map-based cloning in the post-genome era. Plant Physiology 129,440-450.

Kurata N, Nagamura Y, Yamamoto K et al. (1994) A 300 kilobase interval genetic map of rice including 883 expressed sequences. Nature Genetics 8,365-372.

McCallum CM, Comai L, Green EA and Henikoff S (2000) Targetting induced local lesions in genomes (TILLING) for plant functional genomics. Plant Physiology 123, 439-442.

Moncada P, Martinez CP, Borrero J, Chatel M, Gauch H, Guimaraes E, Tohme J and McCouch SR (2001) Quantitative trait loci for yield and yield components in an *Oryza sativa* x *Oryza rufipogon* BC2F2 population evaluated in an upland environment. Theoretical and Applied Genetics 102:41-52.

Paterson AH, Lin YR, Li ZK et al. (1995) Convergent domestications of cereal crops by independent mutations at corresponding genetic loci. Science 269,1714-1718.

Peng J, Richards DE, Hartley NM et al. (1999) 'Green revolution' genes encode mutant gibberellin response modulators. Nature 400,256-261.

Rossberg M, Theres K, Acarkan A et al. (2001) Comparative sequence analysis reveals extensive microcolinearity in the lateral suppressor regions of the tomato, *Arabidopsis*, and *Capsella* genomes. Plant Cell 13,979-988.

SanMiguel PJ, Ramakrishna W, Bennetzen JL, Busso CS and Dubcovsky J (2002) Transposable elements, genes, and recombination in a 215-kb contig from wheat chromosome 5A(m). Functional and Integrative Genomics 2,70-80.

Shimamoto K and Kyozuka J (2002) Rice as a model for comparative genomics of plants. Annual Review of Plant Biology 53,399-419.

Song R, Llaca V and Messing J (2002) Mosaic organization of orthologus sequences in grass genome. Genome Research 12,1549-1555.

Tarchini R, Biddle P, Wineland R, Tingey S and Rafalski A (2000) The complete sequence of 340kb of DNA around the rice *Adh1-Adh2* region reveals interrupted colinearity with maize chromosome 4. Plant Cell 12,381-391.

Terada R, Urawa H, Yoshihsige I, Thugane K and Lida S (2002) Efficient gene targeting by homologous recombination in rice. Nature Biotechnology 20,1030-1034.

The *Arabidopsis* Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408,796-815.

The Rice Full-Length cDNA Consortium (2003) Collection, mapping, and annotation of over 28,000 cDNA clones from *japonica* rice. Science 301,376-379.

Tikhonov AP, Bennetzen JL and Avramova ZV (2000) Structure domains and matrix attachment regions along collinear chromosomal segments of maize and sorghum. Plant Cell 12,249-264.

Vandepoele K and Simillion C (2003) Evidence that rice and other cereals are ancient aneuploids. Plant Cell 15,2192-2202.

Ware D and Stein L (2003) Comparison of genes among cereals. Current Opinion in Plant Biology 6,121-127.

Wilson WA, Harrington SE, Woodman WL, Lee M, Sorrells ME and McCouch SR (1999) Inferences on the genome structure of progenitor maize through comparative analysis of rice, maize and the domesticated panicoids. Genetics 153,453-473.

Xiao JH, Li JM, Grandillo S, Ahn SN, Yuan LP, Tanksley SD and McCouch SR (1998) Identification of trait-improving quantitative trait loci alleles from a wild rice relative, *Oryza rufipogon*. Genetics 150:899-909.

Xu F, Lagudah ES, Moose SP and Riechers DE (2002) Tandemly duplicated safener-induced glutathione *S*-transferase genes from *Triticum tauschii* contribute to genome- and organ-specific expression in hexaploid wheat. Plant Physiology 130,362-373.

Xu Y. 2002. Global view of QTL: rice as a model. In: Kang MS (ed.) Quantitative genetics, genomics and plant breeding. Wallingford (UK): CAB International. p.109-134.

Xu Y (2003) Developing marker-assisted selection strategies for breeding hybrid rice. Plant Breeding Reviews 23,73-174.

Xu Y, Ishii T and McCouch SR (2003) Marker-assisted evaluation of germplasm resources for plant breeding. In: Rice science: innovations and impact for livelihood. International Rice Research Institute.

Yu J, Hu SN, Wang J et al. (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). Science 296:79-92.