PERSPECTIVES: THE RICE GENOME

Opening the Door to Comparative Plant Biology

Jeffrey Bennetzen

Rice belongs to the grass family, which includes maize, wheat, barley, sorghum, and sugarcane. Together these crop plants provide most of the world's food and animal feed. There is great interest in analyzing the genome of rice because this grass has many of the characteristics of a model plant. Unlike most other grasses, rice has a relatively small genome of about 440 Mb (the maize genome is 2500 Mb and that of barley, 4900 Mb) (*1*). Rice researchers have developed important tools for genetic analysis, including excellent genetic maps (*2*) and efficient genetic transformation techniques (*3*). Comparative genetic maps within the grass family indicate extensive regions of conserved gene content and order (*4*). Thus, identification and study of rice genes can provide both clones and valuable information for any researcher investigating similar traits or regions in the genome of another grass (*5*). The release of draft genome sequences of two rice varieties, published on pages 79 (*6*) and 92 (*7*) of this issue, and ongoing efforts to compile a complete rice genome sequence reveal the tremendous value of rice as a model plant.

In 1998, an international consortium led by the Rice Genome Research Program in Tsukuba, Japan, began to sequence the rice genome. The participants in this project took a traditional approach to genome sequencing: stepwise sequence analysis of a minimal tiling path of overlapping clones containing large inserts of rice DNA. Contiguous maps (contigs) were produced from the genomic DNA in large-insert clone libraries (*8*) for the Nipponbare variety of rice. Researchers in individual nations committed to sequencing specific chromosomes or chromosome segments. This sequencing strategy is slow and expensive, but provides the most precise and complete sequence, with a goal of 99.99% accuracy across the entire genome. To date, this project has yielded over 74 Mb of completed sequence, covering more than 15% of the genome (information available at http://rgp.dna.affrc.go.jp).

The author is in the Department of Biological Sciences, Purdue University, West Lafayette, IN 47907, USA. E-mail: maize@bilbo.bio.purdue.edu

Shortly after the initiation of the International Rice Genome Sequencing Project (IRGSP), Monsanto funded Leroy Hood's group at the University of Washington to produce a draft sequence of most of the genome of the Nipponbare variety. The Monsanto approach involved low-redundancy sequencing of a contig of bacterial artificial chromosome (BAC) clones that covered about 260 Mb of the genome. This strategy identified over 95% of the genes in these sequenced BACs, but does not provide enough information for highly accurate sequence assembly. However, because Monsanto did not produce a complete genome sequence, they could finish their project much more quickly and inexpensively than the IRGSP. Monsanto offered these clones and their sequence data to the IRGSP to assist sequencing of the complete rice genome.

An Inexpensive and Quick Strategy

More recently, the Beijing Genomics Institute (BGI) and Syngenta's Torrey Mesa Research Institute (TMRI) independently decided to produce draft sequences of the rice genome by the fastest and least expensive method: shotgun sequence analysis of small-insert clones. Using the Nipponbare variety, Syngenta sequenced enough random clones (just over 5.5 million) so that their final data set provided an approximate sixfold redundancy (6× coverage) for the entire genome. They obtained 99.8% sequence accuracy, and identified over 99% of the genes at a cost of about 10% of the IRGSP's traditional strategy. BGI, a newly established genome research facility in China, and their collaborators chose to generate shotgun sequence data for two different rice cultivars, *93-11* and *PA64s*, which contribute the paternal and maternal genes, respectively, to the highly productive Chinese rice hybrid, *Liang-You-Pei-Jiu*. Both cultivars contain genes from the *indica* subspecies, which provides most of the world's rice; the Nipponbare inbred variety studied by the other programs is a *japonica* variety typical of the short-grain rice preferred in Japan and some other countries. Comparison of the BGI data with other rice genome sequence data will provide insights into rice genetic diversity.

The BGI project provides a wonderful example of the speed and efficiency of shotgun sequencing. This group was only established in late 1999, yet has already produced 4× redundant coverage of the *93-11* cultivar sequence, and 1.1× redundant coverage of the *PA64s* sequence, as well as detailed annotations and other analyses (see the News story by Normile on page 36). It seems likely that the release of TMRI's sequence (assembled in early 2000, but initially only available to the public sector research community with certain restrictions) was motivated by the rapid progress of the BGI program. Now, both the TMRI and BGI projects are providing their data for inspection and use by any interested parties at www.tmri.org and http://btn.genomics.org.cn/rice. Even though Syngenta's TMRI is a for-profit company, they require no reporting or intellectual property commitments from the users of this resource, only a commitment that the data set not be transferred to any third party. This is a standard component of all research information exchanges in the academic sector, and is (in fact) much less obligating than the material transfer agreements that are now routinely required for exchanges between academic scientists.

The gene identification efficiency of the BGI project is about 92% because sequencing has only progressed to 4× redundancy for *93-11* (and because data generated with the second variety cannot be combined with complete confidence). However, the predicted sequence accuracy is better than 99.9% for more than 90% of the assembled sequence. As with the TMRI project, shotgun sequencing alone does not provide locations of the sequenced segments on the genetic or physical rice genome maps. Because the median length of assembled contiguous sequence is less than 7 kb for each project, the produced sequence has tens of thousands of contigs whose genetic locations are unknown. Locating sequenced regions on the genetic and physical maps is essential for any comprehensive use of this resource. The TMRI project has already anchored most of their sequence contigs onto the genetic and physical maps, and BGI plans to do the same (*6*).

Rice Genome Size and Composition

The two shotgun sequence analyses allowed assembly of similar amounts of low copy–number components of the rice nuclear genome, and also of both the chloroplast and mitochondrial organellar genomes. The estimate of genome size with the 4× BGI data (466 Mb) is substantially higher than the TMRI estimate of 420 Mb. The shotgun approach meant that many repetitive sequences in the nuclear DNA could not be assembled; the two groups used somewhat different strategies

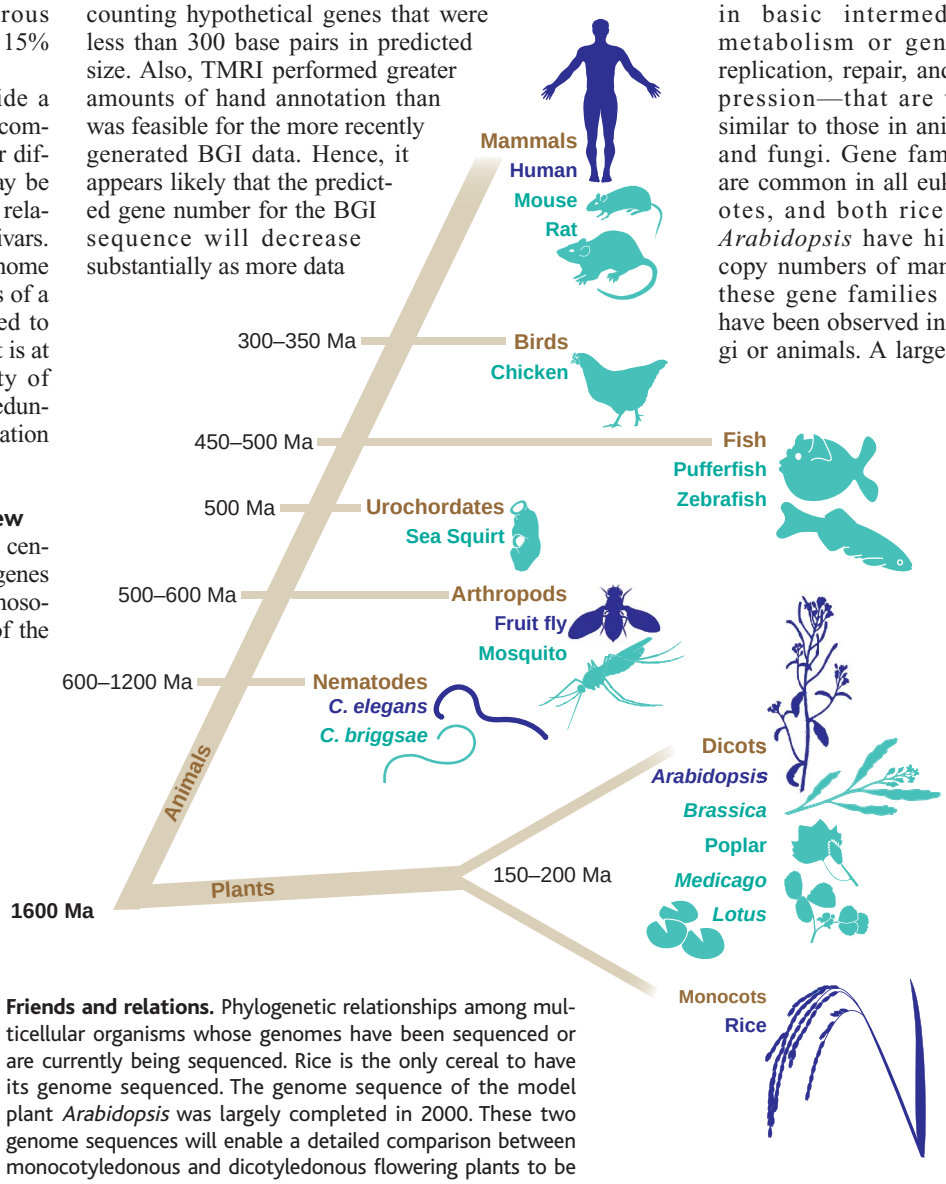
to mask or otherwise set aside many DNA repeats. Despite this limitation, BGI and TMRI came up with similar predictions for the total percent contribution of repetitive DNA to the rice nuclear genome, 42% and 45%, respectively. The most abundant repeats are the miniature inverted-repeat transposable elements (MITEs)—98,000 or more copies of these tiny repeats constitute only about 4% of the rice genome. Retrotransposons, the most numerous large repeats, account for more than 15% of the rice genome in each study.

Taken together, the studies provide a very similar general description of the composition of the rice genome. The minor differences between the two analyses may be due to real differences in the size and relative repeat content of the different cultivars. For instance, differences of 20% in genome size are not unusual within populations of a plant species, and are largely attributed to variations in DNA repeat content (*9*). It is at least equally likely that the majority of these differences reflect the different redundancies, assembly strategies, and annotation approaches of the two projects.

Some Genes are Common, Some New

In any genome-sequencing project, the central goal is the discovery of all of the genes in the target organism (and their chromosomal positions). Perhaps the enormity of the data produced is responsible for the resultant fascination with identifying an absolute "gene number" for the sequenced species, thus providing a minimalist summary of the entire set of discoveries. The observation that the worm *Caenorhabditis elegans* has a higher gene number than the more complex fruitfly, *Drosophila melanogaster*, was unsettling to some; the discovery that the weed *Arabidopsis thaliana* has about as many genes as humans was even more disquieting. Something about human nature apparently requires that, to view ourselves as a superior species, we must have the highest gene number. We will need to get over this, as there will be lots of "lower" species (especially polyploids) that have many more genes than humans. It is interesting that many of the most dramatic scientific discoveries have moved man away from an exalted status: Copernicus's refutation of a geocentric universe, Darwin's prediction that humans arose from a "subhuman" ancestor, and the most recent data showing that humans have an average number of genes for a higher eukaryote.

The annotation of both rice draft sequences places the gene number for this grass at the top of all sequenced organisms so far, with TMRI predicting 33,000 to 50,000 genes, and BGI predicting 53,000 to 65,000 genes. The two predictions of gene number for rice, in addition to being high relative to human (~35,000) or *Arabidopsis* (~25,000), are also quite different. There are several possible reasons for this. The TMRI group was more conservative in their gene annotation, for instance discounting hypothetical genes that were less than 300 base pairs in predicted size. Also, TMRI performed greater amounts of hand annotation than was feasible for the more recently generated BGI data. Hence, it appears likely that the predicted gene number for the BGI sequence will decrease substantially as more data are generated and annotated. However, there will also be attenuation of this shrinkage by addition of genes, especially those that encode small peptides or untranslated small RNAs, which are underrepresented by current annotation approaches. Genome annotation is an imperfect process (to be generous), but it continues to improve at a substantial pace.



Friends and relations. Phylogenetic relationships among multicellular organisms whose genomes have been sequenced or are currently being sequenced. Rice is the only cereal to have its genome sequenced. The genome sequence of the model plant *Arabidopsis* was largely completed in 2000. These two genome sequences will enable a detailed comparison between monocotyledonous and dicotyledonous flowering plants to be made. Species in dark blue are those with completed sequences or drafts that have been published; sequencing of genomes for species in turquoise is ongoing. Ma, millions of years ago.

The types and relative numbers of genes in rice look fairly similar to those in *Arabidopsis*. About one-third of the genes found in these two plant species are not found in any fungal or animal genome sequenced so far. These include the many thousands of genes involved in photosynthesis and photomorphogenesis. As noted previously, plants contain many genes—particularly those involved in basic intermediary metabolism or genome replication, repair, and expression—that are very similar to those in animals and fungi. Gene families are common in all eukaryotes, and both rice and *Arabidopsis* have higher copy numbers of many of these gene families than have been observed in fungi or animals. A large percentage of these duplicated genes are on unlinked chromosomes, among colinear clusters of other duplicated genes. This suggests that rice, like *Arabidopsis*, has undergone numerous episodes of polyploidy and/or segmental duplication. Undoubtedly, many of these genes encode proteins with very different functions in plants and animals; one would also expect

that functional divergence took place in the 150 to 200 million years since the progenitors of *Arabidopsis* and rice diverged from a common ancestor (see the figure).

In both projects, more than 80% of the genes that have been annotated in *Arabidopsis* are also found in rice. Both the BGI and TMRI projects find that more than 45% of their predicted rice genes do not have identified homologs in *Arabidopsis*. In both analyses, most of these "extra" rice genes are those that are the most hypothetical, that is, they have not been found in expressed sequence tag (EST) databases. There is little reason why real rice genes that are expressed at low levels (where ESTs would be rare) should be less conserved in evolution than genes expressed at higher levels. This implies that most of the rice "genes" not found in *Arabidopsis* are actually artifacts of annotation. Ongoing improvements in annotation will undoubtedly bring gene numbers down in both rice and *Arabidopsis*, but should also yield a higher concurrence of genes between the two plant species. Nonetheless, real differences in gene content between rice (a monocot) and *Arabidopsis* (a dicot) may be responsible for the physiological and developmental specificities that differentiate these two important flowering plants.

Mining Information: Added Value

Even with an incomplete draft, the existence of a comprehensive rice genome data set provides a powerful tool for life-science researchers. Evolutionary biologists can mine these sequences to help understand how gene families are created, amplified, and diverge to create new biological activities and specificities. Similar questions can now be answered for the different classes of repetitive DNAs that have come to quantitatively dominate most higher eukaryotic genomes. Pharmacologists, physiologists, developmental biologists, and biochemists can inspect 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. Molecular biologists can use the full set of genes to permit comprehensive characterization of gene expression by any of several high-throughput approaches, such as microarray hybridization. Structural biologists can inspect the complete set of predicted and known peptides to identify those that are of most interest for three-dimensional characterization. Geneticists acquire an almost unlimited number of DNA markers. Quantitative geneticists gain instantaneous access to the genes in their segregating population that may be responsible for the traits they have mapped. Population geneticists can use the identified genes as a starting point for the study of allelic variability and distribution in rice. The initial generation of sequences for two rice subspecies, *japonica* and *indica*, plus a growing data set for a third rice variety (*PA64s*) from the BGI group, provides a tremendous start for future characterizations of linkage disequilibrium and for associative genetics. Comparative geneticists will have unlimited opportunities to relate specific changes in gene structure and content to differences in the evolved biology of different plant species.

As in any breaking field of biology, we often find that our predictions of the outcomes of genome sequence generation and analysis are completely wrong. The gene number controversy provides one excellent example, but so does the observation that many plant DNA replication and repair enzymes are more similar to their human homologs than human enzymes of this type are to their *Drosophila* homologs (*10*). We can expect scores of additional discoveries of this type, reminding us how little we comprehend the nature and evolution of even basic biological processes. More importantly, the genes identified by the sequencing projects now provide the raw material to determine why particular characteristics are shared or not shared by specific lineages of organisms. Hence, any genome sequencing project will be synergistic with all that went before and all that will follow, providing the framework to which the life sciences and, in particular, comparative biology can be tethered.

Identification of genes and characterization of gene variation by genomic sequencing only provides a correlation with a particular biological process. Thus, a completed genome sequence is the first step toward a "candidate gene" approach to biology. Detailed and comprehensive studies of gene function are an essential next step. There are already tools of this type available for rice: expression arrays and mutagenized populations that are essential for reverse or forward genetics (*11*, *12*).

The most urgent next step for rice researchers is to place all of the sequenced gene regions on the physical and genetic maps of rice. There are numerous techniques available for pinpointing the thousands of rice sequence contigs precisely and cost effectively in a few months. This accomplishment will allow all past, present, and future mapping studies to be associated with candidate genes, benefiting any researcher interested in a heritable (mappable) trait. The colinearity and common gene content among the grasses indicates that this information will also enable discovery and sequence assembly/annotation across a wide range of important plant species.

Relatively few plant species have been subjected to comprehensive sequence analysis (see the figure). Comparisons across species are incredibly valuable, but the questions that can be asked are quite different, depending on the degree of relatedness. For instance, identification of conserved processes between plants and animals will help us to understand the very basis of multicellular existence. However, to understand what makes mouse different from human, or rice from *Arabidopsis*, we need to identify both the shared and diverged genic complements of the individual species. Among animal research groups, an informal "rule of pairs" has developed to allow characterization of processes that evolved specifically within a lineage of organisms—hence the sequence analysis of both mouse and rat, pufferfish and zebrafish, fruit fly and mosquito (see the figure). Plants are exceptional in the great breadth of characterized species for which there are advanced genetic tool kits, thus providing an extra incentive to genomic investigations that are informed by phylogenetic considerations. Plant researchers have now begun to fill out their branch of the tree of life with genome-sequencing projects for the cabbage *Brassica oleraceae* (serving as a twin to *Arabidopsis*) and for two legumes, *Lotus japonicus* and *Medicago truncatula*. Notable by their absence are any paired species for rice, or any representatives of branches of the plant kingdom older than flowering plants, such as ferns or conifers. Despite the unrivaled contributions of the cereal grasses to world food production and their premier status among genetic systems in plants, no other cereal genome is currently undergoing genomic sequence analysis. When such projects are initiated, they will benefit greatly from the availability of rice sequences for gene annotation and map assembly. More important, the comparisons and contrasts between the grasses will provide our first step toward understanding the commonalities and specific niche exploitations that have made this family of plants exceptionally successful since it first emerged 50 to 70 million years ago.

References

1. K. Arumuganathan, E. D. Earle, *Plant Mol. Biol. Rep.* **9**, 208 (1991).
2. Y. Harushima *et al.*, *Genetics* **148**, 479 (1998).
3. Y. Hiei *et al.*, *Plant J.* **6**, 271 (1994).
4. M. D. Gale, K. M. Devos, *Science* **282**, 656 (1998).
5. J. L. Bennetzen, M. Freeling, *Genome Res.* **7**, 301 (1997).
6. J. Yu *et al.*, *Science* **296**, 79 (2002).
7. S. A. Goff *et al.*, *Science* **296**, 92 (2002).
8. S. Saji *et al.*, *Genome* **44**, 32 (2001).
9. A. L. Rayburn *et al.*, *Am. J. Bot.* **72**, 1610 (1985).
10. The *Arabidopsis* Genome Initiative, *Nature* **408**, 796 (2000).
11. H. Hirochika, *Plant Mol. Biol.* **35**, 231 (1997).
12. J. Yazaki *et al.*, *DNA Res.* **7**, 367 (2000).