Genome studies and molecular genetics
The rice genome and comparative genomics of higher plants
Editorial overview

Takuji Sasaki and Ronald R Sederoff

Current Opinion in Plant Biology 2003, **6**:97–100

1369-5266/03/$ – see front matter

DOI 10.1016/S1369-5266(03)00018-9

Takuji Sasaki

Genome Research Department, National Institute of Agrobiological Sciences, 1-2 Kannondai 2-chome, Tsukuba, Ibaraki 305-8602, Japan
e-mail: tsasaki@nias.affrc.go.jp

Takuji is director of the Rice Genome Research Program (RGP), which aims to completely sequence the rice genome and subsequently to pursue integrated goals in functional genomics, genome informatics and applied genomics. The RGP is now the leading member of the International Rice Genome Sequencing Project (IRGSP), a consortium of ten countries sharing the sequencing of the 12 rice chromosomes.

Ronald R Sederoff

Forest Biotechnology Box 7247, 2500 Partners II Building, Centennial Campus, NC State University, Raleigh, North Carolina 27695, USA
e-mail: ron_sederoff@ncsu.edu

Ron is Distinguished University Professor of Forestry and the Co-director of Forest Biotechnology at North Carolina State University. He has worked on the molecular genetics of forest trees, and most recently on the genomics of loblolly pine and *Eucalyptus*. Ron's team have been sequencing genes that are involved in wood formation and using microarrays to learn about the basis of wood properties. They are also interested in the sequence relationships of angiosperms and gymnosperms.

Abbreviations
QTL quantitative trait locus
SNP single nucleotide polymorphism

Year 2002 must be remembered as the year of rice genome sequence. In April, a rough draft sequence of *indica* rice variety 93-11 was published by an academic institute, the Beijing Genomics Institute (BGI, China). In the same month, a draft sequence of *japonica* variety 'Nipponbare' was published by a private company, Syngenta. Both draft sequences were obtained by a whole-genome shotgun approach. In November, two members of the International Rice Genome Sequencing Project (IRGSP), namely the Rice Genome Research Program (RGP, Japan) and the National Center for Gene Research (NCGR, China) published high-quality phase-3 sequences of chromosome 1 and chromosome 4, respectively. Both sequences were obtained by a clone-by-clone strategy. On December 18th, the IRGSP announced the completion of a high-quality draft with at least phase-2 sequences of the 12 rice chromosomes.

The IRGSP was organized at a workshop adjunct to the 5th International Congress of Plant Molecular Biology held in Singapore in 1997. At that time, the molecular genetic analysis of rice was mainly focused on the construction of fine genetic maps with RFLP or SSR markers. Meanwhile, the genome sequencing of *Arabidopsis*, a weedy plant from the crucifer family, was already underway as an international collaborative project. Almost all of the participants in the workshop agreed that rice must be the next target for genome sequencing because it is one of the world's most important cereal crops and has the smallest genome of the cereals. Rice could be a good reference plant for cereals because the availability of many genomic tools and extensive knowledge from other fields of rice research.

Delseny (pp. 101–105) reviews the history of the IRGSP and the competition between private companies for the honor of being the first to completely sequence the rice genome. He discusses the distribution of variation in nucleotide composition, repeated sequences, and gene content within the rice genome, and emphasizes the practical importance of obtaining a finished-quality sequence of the entire genome.

A major long-term goal of genomics is the potential ability to predict phenotype from genotype. If we know the genome sequence of an organism, the functions of all of its genes and the effects of specific alleles, then it may be possible to make strong inferences about the resulting phenotype. From an engineering perspective, one would want to design genomes for specific uses, or at least to be able to modify the genome and to predict the resulting

changes in phenotype. At present, it is possible to make such inferences only for a few specific changes in a small number of single genes. It will never be possible, however, to predict the phenotype of an entire organism as complex as a plant. Not only are the numbers of genes too large but the quantitative nature of variation will be too difficult to resolve, and the genome-wide effects of epistasis will greatly limit such predictions. Nonetheless, the underlying rational for identifying and characterizing all of the genes and proteins in a variety of organisms, which is comparative genomics, is the eventual prospect of making predictions of phenotype from genotype. Although it seems impossible for such a strategy to be fully predictive, we have already taken major steps along that road and are unlikely to turn back.

It is necessary to understand the functions common to all higher plants in order to understand the basis for the great morphological, physiological and chemical diversity among the 250 000 known species of seed plants. Comparative genomics can tell us what is essential for the minimal functions of a plant, for its normal growth, development and metabolism. In addition, comparative genomics makes it becomes possible to learn the basis for specialization and diversity, to understand the molecular basis for adaptation, and to identify mechanisms that underlie the extraordinary level of chemical diversity in plants. In order to understand ''how things are'', it is essential that we learn ''how they got to be that way''. Such information is of more than academic interest as we begin to consider more dramatic types of genetic engineering, such as the modification of plants to become far more specialized metabolic factories, or what is needed for plants to survive in extreme terrestrial environments, or even in extraterrestrial ones.

All vascular plants are thought to have evolved from a small leafless, rootless ancestor about 400 million years ago. Within the subsequent 100 million years, the lineages for the major groups of seed plants (gymnosperms and angiosperms) were established and large areas of land were covered by forests, transforming the surface of the land and the ecology of the planet. During the next 100 million years, the major lineages of monocotyledons and dicotyledons appeared and diversified. 100 million years ago, much of the diversity within the groups of flowering plants was established. The expansion and diversification that has taken place in the past 100 million years, particularly since the end of the Cretaceous about 65 million years ago, represents a small part (a fourth to a seventh) of the history of vascular plants. Much of the entire history of vascular plant evolution is reflected in the DNA sequences of extant plants and will be discovered through comparative genomics.

Just as the sequencing of the *Arabidopsis* genome was a major event in the genomics of plants, the announcements of the early drafts of the rice genome in 2002 were landmark events for comparative genomics. The complete sequence of *Arabidopsis* has already made an impact on plant research and much more remains to be learned from *Arabidopsis* through the functional characterization of mutations. Studies of other genomes are required to answer major questions about genome function and genome evolution, and to apply genomic information to practical problems such as increasing the yield and quality of food crops and other plant products.

The most fundamental issue of comparative genomics is the number and variation in the number of functional plant genes. How many functional genes are there in any plant? To what extent do all plants contain the same genes? The number of genes inferred from the *Arabidopsis* genome sequence is about 26 000, whereas the number estimated for rice ranges from about 32 000 to 62 000. A large number of the genes inferred from the sequence have not been identified as transcripts, however, and it is not yet clear how many of the proposed genes are functional. Many may be transcribed rarely or under unusual circumstances (see review by Schoof and Karlowski [pp. 106–112]).

Functional classification, and the definition of gene families, depends fundamentally on the strategy and computational tools used for annotation. Annotation of the rice genome sequence has been more challenging than the annotation of the *Arabidopsis* sequence (see Delseny, and Schoof and Karlowski). Differences in the quality of the sequences and the limitations of the methods used to annotate the first drafts of the rice genome may explain much of the apparent differences between rice and *Arabidopsis*. Although *Arabidopsis* is phylogenetically distant from rice, these two model plants have orthologous proteins that share very similar amino-acid sequences. Hence, the analysis of the *Arabidopsis* sequence provides information that will facilitate the annotation of the rice sequence. Genome information in *Arabidopsis* includes full-length cDNA sequence information and many *Arabidopsis* genes have been isolated and characterized. More full-length cDNAs will be published for rice in the near future and this will also complement the annotation of the rice genome sequence. Annotation of plant genome sequences is becoming increasingly fundamental to future inquiry and applications in plant biology. Annotation appears to be a scientific ''growth industry''.

Most sequences that are found as transcripts and have open reading frames are likely to have essential functions. If one assumes a null mutation rate of 10^{-5} per gene per generation, it is unlikely that sequences will retain open reading frames without maintaining an essential function. Comparisons of the sequences of expressed genes in different plant species should identify functional genes

that are related by decent, even though they may have become paralogs and have divergent functions. A major question to be answered is the extent to which the major groups of higher plants share a common functional genome. 400 million years provides many opportunities for gene duplication, gene loss, lineage-specific divergence, and even horizontal transfer.

To understand genome function, it becomes necessary to understand functional genes as members of gene families. The history of gene families becomes a central question for plant genome evolution because it is both a basis for and a reflection of functional diversity. Much of the information on the evolution of plant genomes will come from examination of the diversity of specific gene families in many different groups of plants. The evolution of gene families may occur in several ways. Duplication and loss of family members is a continuing process that is based on established genetic mechanisms. Changes in regulatory control — that is, in the time, place or level of gene expression — may have dramatic biological consequences even if coding sequence is conserved. In addition, lineage-specific sequence divergence is likely to occur as a result of selection for new functions (neofunctionalization). Duplicate genes encoding sequences that have multiple functions may become specialized and diverge by subfunctionalization, thereby releasing the constraints of epistasis. Evidence on the extent and nature of these processes during higher plant evolution will come from comparative genomics.

Izawa, Takahashi and Yano (pp. 113–120) compare the functional pathways controlling flowering in rice and *Arabidopsis*, a topic of special interest because of the role of flowering in the evolution of the angiosperms. First, a key gene in flowering, *CONSTANS*, was isolated using mutant *Arabidopsis* plants. Thereafter, several related genes were found to explain the phenotypes of other *Arabidopsis* mutants. In rice, genes that are involved in flowering or heading were identified after quantitative trait locus (QTL) analysis of heading date. A *CONSTANS* orthologue was isolated in rice using nearly isogenic lines for each QTL associated with photoperiod sensitivity. The completed rice genome sequence should facilitate the identification of candidate genes at several of the remaining QTL that are associated with the heading date of rice. Information from *Arabidopsis* will assist in this process.

In genetics, as in real estate, location is very important. The chromosomal position of a gene affects its regulation and level of expression. Location on a genetic map is an important technical factor for correlating gene effects with phenotype, and forms the basis of quantitative trait analysis. Genomics would be much simpler if the order of genes were common (syntenic) across major groups of plants. Let us assume a rate for simple chromosome rearrangements in plants of one per million years. If this were so, then there would be a very low probability that any pair of genes would be separated by rearrangement (in any comparison of any two species) during 200 million years of angiosperm evolution. Then, a high level of synteny and microsynteny should be expected. Within the grasses, extensive colinearity among rice, sorghum, maize and wheat was indicated by the genetic mapping of very small number of genes. The frequent absence of synteny at the sequence level within the angiosperms is therefore intriguing (see review by Ware and Stein [pp. 121–127]). Comparisons of the sequence of potentially adjacent genes in grasses and *Arabidopsis* show much less synteny than is predicted by the above calculation, implying that other mechanisms of gene duplication and dispersal are frequent. Nevertheless, the order of a large number of small blocks of genes is conserved in rice and *Arabidopsis*.

The colinearity of genes in rice and other cereals is the subject of the review by Bennetzen and Ma (pp. 128–133). These authors argue that rice provides a useful model for other cereals, but re-emphasize the dynamic nature of regions between genes and the extent of microrearrangements, both of which are not readily observed by recombination mapping. Although the genome sequence data for cereals such as maize and wheat are very limited at present, the construction of databases for comparative genomics has begun. The Gramene database presented by Ware and Stein is not only a comparative genome mapping database but also a community resource for rice genomics information. The construction of databases to allow the further use of cereal sequence data is crucial for progress in linking nucleotide sequence information to phenotype. Genome databases must focus on linking genomics information to phenotype, thereby providing useful tools for advanced breeding using molecular biology.

Han and Xue (pp. 134–138) review the sequence variation between the two cultivated subspecies of rice, the long-grained non-sticky *indica* and the short-grained sticky *japonica*. The former is cultivated in India, Southeast Asia and China, and the latter mainly in Japan. They differ not only in morphological characteristics but also in genome structures. Han and Xue show extensive conservation of microcolinearity and gene content between *indica* and *japonica*, but they also discover significant numbers of rearrangements and polymorphisms when comparing the two genomes. Comparative analysis of the genomes sequences of *indica* and *japonica* rice is necessary to reveal the mechanisms of subspecies-specific phenotypes and adaptations. Comparison of the two genomes will allow the design of new of polymorphic markers, such as single nucleotide polymorphism (SNP) markers, for the precise tagging of target phenotypes in breeding and for effective transfer of favorable genes

from one ecotype to another. Although there exist many rice varieties all over the world, farmers and consumers prefer local elite varieties that are closely related to each other. SNP markers are very useful for the detection of differences among such closely related genotypes. Such sequence comparison must be linked to phenotypic variation, which might be caused by natural variation in the sequence of the corresponding gene or its regulating gene(s).

Comparative studies will also provide more information on the evolution of intergenic regions. Studies of intergenic regions in maize (see review by Bennetzen and Ma) first showed the dynamic nature of the intergenic regions and the role of retrotransposons and other elements. These findings provided support for the idea that genome size depends more on the number and activity of mobile elements than on the number of functional genes. Sequencing DNA from any specific genome becomes increasingly challenging as the number of repeated sequence elements within the genome increases. Centromeres and large blocks of simple sequences are important elements of chromosome function and evolution, but represent major challenges for the sequencing of very large genomes.

A related question is the distribution of genes within plant genomes. *Arabidopsis* is characterized by a relatively homogeneous distribution of genes on chromosomes, whereas rice is more gene dense in distal regions of its chromosomes. Ware and Stein compare genes among cereals and suggest a mosaic structure of gene-rich islands that are separated by high-copy DNA. Sequencing any of the very large genomes of gymnosperms might be made easier if gene-rich regions could be identified. Methods for the fractionation of active chromosome regions based on low methylation may become useful in this regard.

Studies of *Arabidopsis* will not provide information on the process of domestication nor will they provide a guide to the production of improved cultivars through selective breeding. It is speculated that the ancestral plant of the grass family appeared 60 million years ago, and eventually diverged into many species. Cultivated rice is believed to originate from a wild relative, *O. rufipogon*. Many of the details of this evolutionary process are unknown but are hidden within the genome sequences of the *Oryza* species and await discovery by comparative genomics. The diversity of the genus *Oryza* is the subject of the paper by Vaughan, Morishima and Kadowaki (pp. 139–146), who chart a path from the rice genome to an understanding of the processes of domestication, speciation, polyploidy and ecological adaptation. They report that 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. For example, wild *Oryza* species are a potential source of genes for resistance to disease and other stresses, which could allow the cultivation of rice in less favorable conditions and increase rice yields.

The history of agriculture, and genetic studies of maize, argues strongly that small numbers of genetic changes have provided major steps in the domestication of many species. Equally important, has been selection for quantitative variation in plant breeding. Genomics has provided new approaches to investigate and utilize both qualitative and quantitative variation, which should prove useful for the next generation of plant breeding. The detection of SNPs that are associated with expressed sequence tags (ESTs) will identify sequence variation that can be associated with variation in phenotype, but making such associations is not efficient on a genomic scale with large numbers of candidate genes. The assay of variation in gene expression on microarrays promises to be a powerful approach to determining the effect of specific genes and their SNPs on phenotype.

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 the future. This section focuses on the impact of the rice genome sequence in various research fields. Enjoy.