106. R. J. Cho *et al.*, *Nature Genet.* **23**, 203 (1999).
107. E. Drenkard *et al.*, *Plant Physiol.* **124**, 1483 (2000).
108. J. Hanke *et al.*, *Trends Genet.* **15**, 389 (1999).
109. A. A. Mironov, J. W. Fickett, M. S. Gelfand, *Genome Res.* **9**, 1288 (1999).
110. L. Croft *et al.*, *Nature Genet.* **24**, 340 (2000).
111. D. Brett *et al.*, *FEBS Lett.* **474**, 83 (2000).
112. B. Modrek, A. Resch, C. Grasso, C. Lee, *Nucleic Acids Res.* **29**, 2850 (2001).
113. S. M. Berget, *J. Biol. Chem.* **270**, 2411 (1995).
114. B. J. Blencowe, *Trends Biochem. Sci.* **25**, 106 (2000).
115. M. L. Hastings, A. R. Krainer, *Curr. Opin. Cell Biol.* **13**, 302 (2001).
116. Z. J. Lorkovic, D. A. Wieczorek Kirk, M. H. Lambermon, W. Filipowicz, *Trends Plant Sci.* **5**, 160 (2000).
117. D. Brett *et al.*, *Nature Genet.* **30**, 29 (2002).
118. A. P. Bird, *Trends Genet.* **11**, 94 (1995).
119. Source for BAC-end sequences: http://www.genome.clemson.edu/projects/rice/fpc.
120. We are indebted to faculty and staff at the Beijing Genomics Institute, whose names were not listed, but who also contributed to the team effort (www.genomics.org.cn). We are indebted to our scientific advisors, M. V. Olson, L. Bolund, R. Waterston, E. Lander, and M.-C. King, for their long-term support. We are grateful to R. Wu and C. Herlache for editorial assistance on the manuscript. We thank Amersham Pharmacia Biotech (China) Ltd., SUN Microsystems (China) Inc., and Dawning Computer Corp. for their support and service. This project was jointly sponsored by the Chinese Academy of Science, the Commission for Economy Planning, the Ministry of Science and Technology, the Zhejiang Provincial Government, the Hangzhou Municipal Government, the Beijing Municipal Government, and the National Natural Science Foundation of China. The analysis was supported in part by the National Institute of Environmental Health Sciences (grant 1 RO1 ES09909).

14 November 2001; accepted 20 February 2002

# A Draft Sequence of the Rice Genome (*Oryza sativa* L. ssp. *japonica*)

**Stephen A. Goff,[1]* Darrell Ricke,[1] Tien-Hung Lan,[1] Gernot Presting,[1] Ronglin Wang,[1] Molly Dunn,[1] Jane Glazebrook,[1] Allen Sessions,[1] Paul Oeller,[1] Hemant Varma,[1] David Hadley,[1] Don Hutchison,[1] Chris Martin,[1] Fumiaki Katagiri,[1] B. Markus Lange,[1] Todd Moughamer,[1] Yu Xia,[1] Paul Budworth,[1] Jingping Zhong,[1] Trini Miguel,[1] Uta Paszkowski,[1] Shiping Zhang,[1] Michelle Colbert,[1] Wei-lin Sun,[1] Lili Chen,[1] Bret Cooper,[1] Sylvia Park,[1] Todd Charles Wood,[2] Long Mao,[3] Peter Quail,[4] Rod Wing,[5] Ralph Dean,[5] Yeisoo Yu,[5] Andrey Zharkikh,[6] Richard Shen,[6]† Sudhir Sahasrabudhe,[6] Alun Thomas,[6] Rob Cannings,[6] Alexander Gutin,[6] Dmitry Pruss,[6] Julia Reid,[6] Sean Tavtigian,[6] Jeff Mitchell,[6] Glenn Eldredge,[6] Terri Scholl,[6] Rose Mary Miller,[6] Satish Bhatnagar,[6] Nils Adey,[6] Todd Rubano,[6]† Nadeem Tusneem,[6] Rosann Robinson,[6] Jane Feldhaus,[6] Teresita Macalma,[6] Arnold Oliphant,[6]† Steven Briggs[1]**

The genome of the japonica subspecies of rice, an important cereal and model monocot, was sequenced and assembled by whole-genome shotgun sequencing. The assembled sequence covers 93% of the 420-megabase genome. Gene predictions on the assembled sequence suggest that the genome contains 32,000 to 50,000 genes. Homologs of 98% of the known maize, wheat, and barley proteins are found in rice. Synteny and gene homology between rice and the other cereal genomes are extensive, whereas synteny with *Arabidopsis* is limited. Assignment of candidate rice orthologs to *Arabidopsis* genes is possible in many cases. The rice genome sequence provides a foundation for the improvement of cereals, our most important crops.

Cereal crops constitute more than 60% of total worldwide agricultural production (*1*), and rice, wheat, and maize are the three most important cereals. More than 500 million tons of each are produced annually worldwide; per capita consumption averages as high as 1.5 kg per day (*2*). Most rice grown is consumed directly by humans, and about one-third of the population depends on rice for more than 50% of caloric intake (*3*).

The cereals have been evolving independently from a common ancestral species for 50 to 70 million years (*4*), but despite this long period of independent evolution, cereal genes and genomes display high conservation. Comparisons of the physical and genetic maps of the grass genomes show conservation of gene order and orientation, or synteny (*5–7*). Despite gene similarity and genome synteny, cereal genome sizes vary considerably. The genomes of sorghum, maize, barley, and wheat are estimated at 1000, 3000, 5000, and 16,000 megabase pairs (Mbp), respectively. Rice has a much smaller genome, estimated at 420 Mbp. The small genome and predicted high gene density of rice make it an attractive target for cereal gene discovery efforts and genome sequence analysis.

Over the past several years, selected regions of the japonica and indica rice genomes have been sequenced. The International Rice Genome Sequencing Project (IRGSP) was organized to achieve >99.99% accurate sequence using a mapped clone sequencing strategy (*8*). In addition, expressed gene sequencing has been actively pursued. More than 104,000 expressed sequence tags (ESTs) from a variety of rice tissues have been entered into the EST database (*9*). Other rice genome sequencing projects have been reported by Monsanto Co. (*10*) and by the Beijing Genomics Institute (*11*).

The two major groups of flowering plants, monocots and dicots, diverged 200 million years ago (*12*). In late 2000, the 125-Mbp genome of the dicot model plant *Arabidopsis thaliana* was reported (*13–15*). Similar high-accuracy sequencing projects of important cereals would be expensive and slow because their genomes are so large. Recent improvements in automated DNA sequencing have made whole-genome shotgun sequencing an attractive approach for gene discovery in both small and large genomes (*16–18*). Here, we describe the random-fragment shotgun sequencing of *Oryza sativa* L. ssp. *japonica* (cv. Nipponbare) to discover rice genes, molecular markers for breeding, and mapped sequences for the association of candidate genes and the traits they control. Also reported are the linkages of sequence assemblies to rice bacterial artificial chromosome (BAC) end sequences and fingerprints (*19–22*), anchoring of the physical and genetic maps, and the syntenic relationship between rice and other plants. The finding that most cereal genes have strong rice homologs suggests that the rice genome will be useful as a foundation for sequencing the genomes of

[1]Torrey Mesa Research Institute, Syngenta, 3115 Merryfield Row, San Diego, CA 92121, USA (www.tmri.org). [2]Bryan College, Dayton, TN 37321, USA. [3]Department of Biological Sciences, Northern Illinois University, DeKalb, IL 60115, USA. [4]Department of Plant and Microbial Biology, University of California, Berkeley, CA 94720, USA. [5]Clemson University Genomics Institute, 100 Jordan Hall, Clemson, SC 29630, USA. [6]Myriad Genetics, 320 Wakara Way, Salt Lake City, UT 84108, USA.

*To whom correspondence should be addressed. E-mail: stephen.goff@syngenta.com
†Present address: Illumina Inc., 9885 Towne Centre Drive, San Diego, CA 92121, USA.

related cereals. Synteny among cereals should allow placement of low-copy cereal genes on a rice genome framework.

**Sequence generation, assembly, coverage analysis, and repeats.** A shotgun library of the rice genome was constructed in a low-copy plasmid from purified, sheared nuclear DNA. Clones were sequenced from both ends as described (Web site link 1; for this and other supplementary Web data, see *Science* Online at www.sciencemag.org/cgi/content/full/296/5565/92/DC1). More than 2.5 billion bases with a 98% probability of being correct were generated, representing more than sixfold coverage of the estimated 420-Mbp rice genome (*23*). About 80% of the sequencing reads were linked to a second sequence generated from the opposite end of the same template. After removal of an estimated 38 Mbp of repetitive DNA, more than 5.5 million sequences assembled into 42,109 contiguous sequences (contigs) with a total coverage of 389,809,244 bp (93% of the predicted 420-Mbp rice genome) and a GC content of 44%. This sequence assembly will be referred to as Syd (Syngenta draft sequence; data access information is available at www.tmri.org).

Contigs of nonrice origin were identified by sequence homology to known bacteria, high GC content, lack of homology to rice BAC end sequences, and/or depth of coverage. Sequence analysis identified 6 Mbp as originating from two related bacterial species (*Xanthomonadales*), likely representing endophytes present in the plant material used for DNA isolation. These sequences were not included in Syd; however, some bacterial contigs may remain (Web site link 2).

Syd data were compared to almost 1 million bases of IRGSP's completed rice genome sequence to determine coverage and quality. Syd coverage ranged from 98.7 to 99.8% on different clones (Table 1). Sixty-three gaps were found, totaling 6400 bp or 0.6%; the average gap size was 101 bp. Most of the Syd gaps were the result of conservative assembly; more than 70% of the gaps could be closed by manual editing. A single base pair difference and an insertion-deletion difference (indel) were found once every 1000 and 2000 bp, respectively; these findings indicate that Syd is 99.8% accurate. Gene coverage was assessed by comparison of 495 full-length rice genes with Syd. Of 648,792 bp, 643,528 bp (99.2%) were found in Syd (Web site link 3). Most genes were represented at more than 90% of their full length. The only three genes not found in Syd were determined to be misannotated. The analysis of coverage and quality indicates that Syd contains the overwhelming majority of rice genes, although some gene sequences are partial and/or in more than one contig. Unedited Syd sequence was used for the analysis presented below.

The rice genome sequence is available at www.tmri.org. Copies of the agreements governing access to the data are available on *Science* Online (www.sciencemag.org/cgi/content/full/296/5565/92/DC1) and at www.tmri.org. A summary of the agreements is available at the *Science* Online URL.

Simple sequence repeats (SSRs) are highly polymorphic sequences found throughout plant and animal genomes. SSR repeat unit lengths are easily detectable and have made SSRs a popular type of codominant molecular marker for accelerated breeding. Syd analysis revealed a total of 48,351 dinucleotide SSRs (eight repeat units minimum), trinucleotide SSRs (five repeat units minimum), and tetranucleotide SSRs (four repeat units minimum), or about one SSR every 8000 bp (Table 2). Di-, tri-, and tetranucleotide SSRs account for 24%, 59%, and 17%, respectively, of the SSRs found in rice. The frequency of specific SSRs is not random nor representative of the genome GC content. The most frequent dinucleotide SSR is AG/CT and variants, representing 58% of all dinucleotides. The most frequent trinucleotide is CGG/CCG and variants, at 44% of all trinucleotide SSRs. ATCG/CGAT is the most common tetranucleotide repeat unit (Web site link 4). More than 7000 SSRs were found in predicted genes. Most of these SSRs (92%) are trinucleotides, so length changes should maintain the open reading frame. In addition to SSRs, ~38 Mbp of long repetitive DNA and 150 Mbp of short repetitive DNA were identified. A detailed analysis of the repetitive DNA, retrotransposons, organellar DNA, rDNA, and tRNA genes, as well as miniature inverted-repeat transposable elements (MITEs) in the rice genome, can be found at Web site link 5.

**Prediction and classification of Syd genes.** Gene prediction in silico is an imperfect process (*24*), and no single gene-prediction algorithm was found to be highly accurate on Syd; therefore, several approaches were combined to identify genes. Five different combinations of gene prediction programs and training models were used to identify genes (Web site link 6). Additional genes were predicted on the basis of homology to plant and fungal genes. Timelogic's Decypher FrameSearch algorithm was used to detect and guide the correction of frameshifts caused by indels. For each predicted gene, the fraction of the length with homology to known genes, predicted genes from other species, Prosite motifs (*25*), or Pfam domains (*26*) was used as a confidence score. When predicted genes overlapped, the one with the highest confidence score was selected. Predicted genes were separated into three categories: high (*H*genes), with confidence scores of ≥75%; medium (*M*genes), with confidence scores from 1 to 75%; and low (*L*genes), with confidence scores of <1%. Predicted genes were often found to be incomplete, lacking either an $NH_2$-terminal or COOH-terminal coding region (Web site link 6).

The number of genes identified in Syd depends on the minimum gene length chosen. More than 78% of *L*genes are shorter than 500 bp, whereas only 42% of *M*genes and 28% of *H*genes are shorter than 500 bp. Including *H*genes and *M*genes longer than 500 bp and *L*genes longer than 1000 bp yields

**Table 1.** Genome coverage in Syd. Six fully sequenced rice BACs (GenBank) were compared to sequences in Syd to determine coverage and analyze gaps. Sequences were aligned and gaps analyzed for length and number. Manual gap closure was performed on contigs covering two BACs, and the analysis was repeated.

| BAC | Length | Gaps | Gap length | Coverage (%) |
|---|---|---|---|---|
| BAC1 | 156,054 | 14 | 1992 | 98.7 |
| BAC2 | 183,173 | 7 | 617 | 99.7 |
| BAC3 | 170,777 | 7 | 333 | 99.8 |
| BAC4 | 156,067 | 11 | 743 | 99.5 |
| BAC5 | 154,555 | 11 | 896 | 99.4 |
| BAC6 | 157,998 | 13 | 1,789 | 98.8 |
| Total | 978,624 | 63 | 6370 | — |
| Edited to close gaps: | | | | |
| BAC1 | 156,054 | 3 | 1100 | 99.3 |
| BAC2 | 183,173 | 3 | 550 | 99.7 |

**Table 2.** Di-, tri-, and tetranucleotide SSRs in Syd. Simple sequence repeats in Syd were identified computationally and classified as di-, tri-, or tetranucleotide SSRs. The most frequent SSR types do not include base-shifted variants (e.g., the frequency for CGG/CCG does not include GGC/GCC and GCG/CGC SSRs). For distributions with variants, see Web site link 4.

| SSR | Occurrence | *n* | Most frequent type | Observed/expected |
|---|---|---|---|---|
| $(NN)_n$ | 11,416 | 8 to 77 | AG/CT | 3013/1902 |
| $(NNN)_n$ | 28,647 | 5 to 39 | CGG/CCG | 6373/954 |
| $(NNNN)_n$ | 8,288 | 4 to 60 | CGAT | 407/65 |
| Total | 48,351 | — | — | — |

17,164 *H*genes, 12,030 *M*genes, and 3083 *L*genes, for a total estimate of 32,277 genes (Web site link 6). Alternatively, if the minimum gene length for *L*genes is also set at 500 bp, then the total estimate is 40,884 genes. The more inclusive set of 61,668 *HML*genes longer than 300 bp (*HML*genes$^{300}$) was used for the rest of this study, unless noted otherwise. Determining the definitive number of rice genes will require substantial efforts in functional genomics.

Translated *HML*genes$^{300}$ were classified with the software package INTERPRO (*27, 28*). INTERPRO output was filtered to create sets of the longest protein domain for each associated protein, and domains were categorized using Gene Ontology (GO) software (*29*). The results of these classifications are shown in Fig. 1; about 44% of *H*genes, 32% of *M*genes, and 5% of *L*genes were classified, respectively. Most of the classified proteins fall into the categories of metabolism and cell communication/signal transduction.

**Gene and chromosomal duplications.** Global duplication of predicted genes was determined by comparison (using BLAST) of all *H*genes and *M*genes (37,777) longer than 300 bp against one another. Of these, 77% (29,226) were found to be homologous to at least one other predicted gene (TBLASTX E value ≤ –20). *HM*genes$^{300}$ comprise about 15,000 distinct gene families, similar to the 11,000 to 15,000 families predicted for *Arabidopsis* (*13*), *Caenorhabditis elegans* (*30*), and *Drosophila* (*16*). Local duplications were determined by comparing *H*genes$^{300}$ and *M*genes$^{300}$ that were mapped to a single BAC contig. A total of 791 BAC contigs, averaging 500,000 bp, contained 25,728 genes. The fraction of locally duplicated genes ranged from 15.4 to 30.4%, depending on the chromosome (Table 3).

Chromosomal duplications were identified by comparison (BLASTN) of more than 2000 mapped rice cDNA markers (*31*) to the anchored portion of Syd. At a minimum of 80% identity over 100 bp, 851 markers (41%) were single loci, 509 (24%) had two copies, and the remainder (35%) had three or more copies. Duplications were plotted by chromosomal region (Web site link 7). The smallest conserved evolutionary units (SCEUS) method (*32, 33*) was applied to determine the extent of genome duplication. Most SCEUS carry four or fewer mapped markers, suggesting extensive recombination/rearrangement since the duplication events.

The amino acid substitution rate ($d_A$) was used to estimate the age of genome duplications (*34–37*). Whereas a maize whole-genome duplication is reported to have occurred 11.4 million years ago (*38*), an apparent rice whole-genome duplication occurred 40 to 50 million years ago (Web site link 8). The largest chromosomal duplication is on chromosomes 11 and 12 (*31, 39*). In an effort to estimate the age of this duplication, proteins on chromosome 11 were compared to their homologs on chromosome 12. The distribution of $d_A$'s indicates a major duplication about 25 million years ago (Web site link 9) (*34*).

**Syd genes compared to *Arabidopsis* genes.** Eighty-five percent of *Arabidopsis* predicted proteins (21,590 of 25,554) were significantly homologous to *HML*genes$^{300}$ predicted proteins; of these, 2565 show very strong conservation between *Arabidopsis* and rice (Fig. 2). The overall mean identity of *Arabidopsis* proteins to rice contigs is 49.5%, with a mode of 33% (Web site link 10). About 30% of the highly conserved genes are classified as "hypothetical," "unknown," or "putative" (*13*), which suggests that many important plant proteins remain completely uncharacterized. One-third (~8000) of *Arabidopsis* predicted proteins are found in rice, but not in *Drosophila*, *C. elegans*, *Saccharomyces*, or sequenced bacterial genomes. These genes are likely to represent the plant-specific set. About 4000 predicted *Arabidopsis* proteins lack significant homology to *HML*genes$^{300}$ or to Syd. Most of these are classified as "hypothetical" or "unknown" and are not found in genome sequences of other organisms, which suggests that some may be inaccurately predicted and others could be dicot-specific. Homologs of more than 13,000 *HML*genes$^{300}$ are not found in other nonplant sequenced genomes but are found in the *Arabidopsis* genome. About 3886 *H*genes$^{300}$ and 31,387 *HML*genes$^{300}$ are not found with significant homology (BLASTP E value ≤ –6) in the *Arabidopsis* genome, but most of these are low-evidence predictions.

The *Arabidopsis* genome is reported to lack several classes of genes found in other sequenced eukaryotic genomes (*13*). These gene classes are also not found in Syd, nor are members of the following families: nuclear steroid receptor, Janus kinase (JAK)/signal transducers and activators of transcription (STAT), Notch/Lin12, transforming growth factor–β/SMAD, Rel, forkhead/winged helix, POU, IF, Wingless/Ent, caspase, p53, and Hedgehog. Specific gene classes are overrepresented in rice and *Arabidopsis*. For example, RING zinc finger proteins and F-box domain proteins are overrepresented in the *Arabidopsis* genome relative to yeast, *Drosophila*, and *C. elegans*

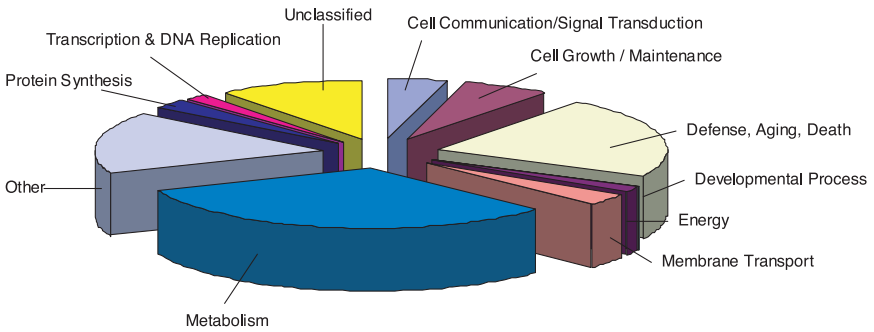



**Fig. 1.** Rice gene prediction classifications. *HML*genes$^{300}$ were classified with Interpro and GO software (*27–29*); the categories generated are shown.

**Table 3.** Local duplication of *H*genes$^{300}$ and *M*genes$^{300}$ in Syd. Genes from individual BAC contigs were compared to each other (TBLASTX), and pairs with E values ≤ –20 were defined as paralogs representing locally duplicated genes.

| Chromosome | BAC contigs | Genes | Paralogs | Average number of genes per contig | Average number of paralogs per contig |
|---|---|---|---|---|---|
| 1 | 99 | 4,467 | 956 (21.4%) | 45 | 9.6 |
| 2 | 74 | 3,011 | 616 (20.5%) | 41 | 8.3 |
| 3 | 89 | 3,197 | 493 (15.4%) | 36 | 5.5 |
| 4 | 61 | 2,679 | 689 (25.7%) | 44 | 11.3 |
| 5 | 68 | 2,426 | 472 (19.5%) | 36 | 6.9 |
| 6 | 67 | 2,342 | 484 (20.7%) | 35 | 7.2 |
| 7 | 69 | 2,507 | 568 (22.7%) | 36 | 8.2 |
| 8 | 65 | 2,286 | 489 (21.4%) | 35 | 7.5 |
| 9 | 47 | 1,618 | 323 (20.0%) | 34 | 6.9 |
| 10 | 55 | 1,724 | 433 (25.1%) | 31 | 7.9 |
| 11 | 48 | 1,834 | 557 (30.4%) | 38 | 11.6 |
| 12 | 49 | 1,870 | 497 (26.6%) | 38 | 10.1 |
| Total | 791 | 29,961* | 6577 (22.0%) | 38 | 8.3 |

*As a result of genome duplication, some sequence contigs (and genes therein) are mapped to multiple locations, thus inflating the total number of mapped genes.

(*13*). These proteins are involved in intracellular protein degradation pathways and their regulation. RING zinc finger and F-box domain proteins are also overrepresented in Syd. More than 840 predicted proteins in *HML*genes$^{300}$ were found to contain RING zinc finger domains, and 150 F-box–containing proteins were identified. This finding is consistent with the speculation that protein turnover and the regulation of protein degradation in plants play an important role in the maintenance of plant homeostasis.

The small phytochrome gene family is underrepresented in rice relative to *Arabidopsis*. Syd contains only three phytochromes (phyA, phyB, and phyC; Web site link 11) of the five genes (phyA to phyE) found in *Arabidopsis*; this finding confirms previous work demonstrating that grasses contain a subset of the five phytochromes found in dicotyledonous plants (*40*). The absence of phyD and phyE, which are partially redundant in function with phyB in *Arabidopsis*, has intriguing evolutionary and regulatory implications, and it suggests that increased gene number in rice relative to *Arabidopsis* is not uniform across the genome.

**Synteny between rice and *Arabidopsis*.** Rice and *Arabidopsis* diverged from a common ancestor about 200 million years ago (*12*). Although the existence of *Arabidopsis*-rice synteny has been controversial (*41, 42*), evolutionary models based on estimated mutation rate predict some syntenic relationships between distantly related species such as *Arabidopsis* and rice (*43*). To address this issue, all annotated *Arabidopsis* proteins were compared to anchored Syd sequence contigs. This approach links *Arabidopsis* proteins to related mapped rice sequences, forming syntenic groups (*44*). When a 99.9% significance threshold is applied, 137 *Arabidopsis*-rice syntenic groups are found at 75 rice chromosomal locations (Table 4) (*44*) throughout the genome, with no discernible pattern. This is a conservative estimate; reducing the significance threshold to 99% increases the number of syntenic groups to 508 (Table 4). Of the 137 high-confidence syntenic groups, the largest mapped to *Arabidopsis* chromosome 5 (from 0 to 26 Mbp) and rice chromosome 4 (from 116 to 129 cM). This syntenic block includes 119 *Arabidopsis* proteins. The predicted roles of these proteins do not suggest an obvious reason for their conservation.

Within the 137 high-confidence syntenic groups, several rice blocks map to more than one site in the *Arabidopsis* genome. One such block maps to all five *Arabidopsis* chromosomes, 8 map to four *Arabidopsis* chromosomes, 10 map to three *Arabidopsis* chromosomes, 14 map to two *Arabidopsis* chromosomes, and the remaining 42 map to a single *Arabidopsis* chromosome. This observation suggests that multiple rounds of duplication occurred within the *Arabidopsis* genome, and it is consistent with the results of studies comparing distantly related dicot pairs such as tomato-*Arabidopsis* (*45*) and soybean-*Arabidopsis* (*46*).

Syntenic protein pairs are two proteins found in close proximity in both rice and *Arabidopsis*, excluding tandem duplications. Only 2% of the syntenic protein pairs on *Arabidopsis* chromosome 5 are adjacent to one another (Web site link 12). Fifty-two percent of the syntenic protein pairs are separated by 1 to 150 intervening proteins. This distribution of related protein pairs in rice and *Arabidopsis* is not random, providing further support for a syntenic relationship between *Arabidopsis* and rice. Selective gene loss and large-scale chromosomal duplication during *Arabidopsis* genome evolution (*45*) could be responsible for the distribution observed.

These observations support previous hypotheses that detectable synteny exists between monocots and dicots even after 200 million years of divergence, although the conservation is less extensive than previously predicted (*43*). The rice and *Arabidopsis* genomes are rearranged to such an extent that constructing a monocot-dicot comparative framework based on these two genomes would be difficult. The low but detectable synteny between rice and *Arabidopsis* may provide clues to orthologous gene identification in future functional genomics studies.

**Genes involved in disease resistance: Conservation between *Arabidopsis* and rice.** Disease resistance genes (*R* genes) are responsible for early and specific recognition of pathogen attack and initiation of signal transduction, leading to deployment of defense mechanisms (*47*). *R* genes fall into two major and three minor structural classes. The largest class of known *R* gene products contains characteristic nucleotide binding (NB) sites, leucine-rich repeats (LRRs), and an apoptosis-resistance-conserved (ARC) domain. The rice genome has ~600 genes that encode proteins with clear NB-ARC homology (*48*). In dicots, NB-LRR genes that encode TIR (Toll–interleukin 1 receptor resistance) motifs at their $NH_2$-termini are very common. For example, among 128 NB-LRR genes in the *Arabidopsis* genome, 85 belong to the TIR-NB-LRR subclass (*13*). In contrast, the rice genome lacks obvious TIR-encoding genes. The regions $NH_2$-terminal to the NB-ARC domains in five predicted rice NB-LRR proteins have very weak homologies to TIR. Possibly these domains function like TIR domains in dicots but are highly diverged. Clearly, TIR-NB-LRR genes are not a major class of *R* genes in rice. This class likely evolved after the divergence of monocots and dicots.

*R* genes encoding extracellular LRRs and either short cytoplasmic tails or COOH-terminal serine-threonine protein kinase domains constitute another major class. Rice has ~450 extracellular LRR genes, and about half encode COOH-terminal protein kinase domains. In contrast to NB-LRR genes, this structural class is known to include proteins with functions unrelated to disease resistance (*47*). Minor classes of *R* genes include those encoding the cytoplasmic serine-threonine kinases Pto and PBS1, each of which have 14 rice homologs; Hs1$^{pro-1}$, which has one rice homolog; and RPW8, which has one rice homolog.

Several plant genes controlling disease resistance signal transduction cascades are known, mainly from *Arabidopsis* (*49*). The site of action of these genes in the signal transduction network is presented at Web site link 13. One rice homolog was found for each of the *Arabidopsis* disease signal transduction genes *NDR1*, *PAD4*, and *EDS1*, as well as the barley gene *RAR1*. Three rice homologs were found for *COI1*, a gene required for responses to the signal

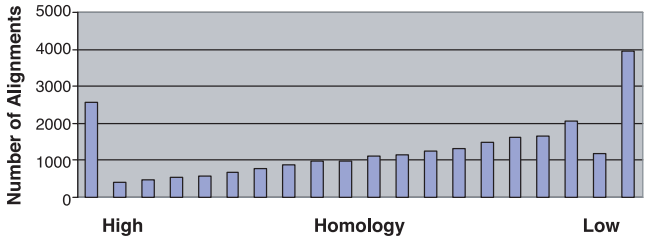



**Fig. 2.** Similarity of 25,554 *Arabidopsis* proteins and best rice homologs. Predicted *Arabidopsis* proteins (October 2001, ftp.tigr.org) were compared (BLASTP E value $\leq -6$) with *HML*genes$^{300}$ translations. The expectation values range from E $< -180$ (high homology) to E $> -6$ (low homology) and are depicted in intervals spanning 10 exponents (e.g., $< -180$, $-180$ to $-171$, $-170$ to $-161$, etc.).

**Table 4.** Rice-*Arabidopsis* synteny. Rice and *Arabidopsis* syntenic groups were detected at various significance thresholds (*44*).

| Chromosome | Significance threshold | | |
|---|---|---|---|
| | 99.9% | 99% | 95% |
| 1 | 41 | 132 | 272 |
| 2 | 34 | 106 | 217 |
| 3 | 31 | 105 | 226 |
| 4 | 11 | 73 | 223 |
| 5 | 20 | 92 | 224 |
| Total | 137 | 508 | 1166 |

molecule jasmonic acid; six for *NPR1*, a gene required for responses to the signal molecule salicylic acid (one closely related and five distantly related); and six for *LSD1*, a gene required for control of programmed cell death (Web site link 14). Many rice homologs for the *Arabidopsis* mitogen-activated protein kinase (MAPK) gene *MPK4* and the MAPK kinase kinase gene *EDR1* were identified, preventing the assignment of putative orthologs. No rice gene similar to *Arabidopsis SNI1* was found. Nearly all genes known to control disease resistance responses in *Arabidopsis* have putative orthologs in rice, suggesting extensive conservation of disease resistance signaling between monocots and dicots.

**Flowering-time and flower development genes.** Flowering in *Arabidopsis* is initiated by flowering-time genes that activate floral meristem identity genes, leading to the patterned expression of floral organ identity genes (Fig. 3). Rice contains single-copy homologs of the *Arabidopsis* flowering-time genes *GI*, *CO*, *LD*, and *FCA* (Web site link 15). The rice *CO*-like flowering-time gene *Hd1* (*50*) is more similar to the uncharacterized *Arabidopsis CO*-related gene *At3g02380* than to *CO*. Tandemly duplicated *FRI* homologs exist in rice, and they are more similar to the *Arabidopsis FRI*-related gene *At5g48390* than to *FRI*. This suggests that *FRI* homologs in rice and *Arabidopsis* play distinct roles not necessarily restricted to the vernalization response. Four rice homologs of the *Arabidopsis GA1* gene that encodes a rate-limiting step in GA biosynthesis were identified (Web site link 16). The *FT/TFL* gene family encodes proteins with homology to phosphatidylethanolamine binding proteins that have been shown to be involved in major aspects of whole-plant architecture (*51, 52*). The rice genome contains 17 members of the *FT/TFL* gene family; one member is most similar to *TFL*, and nine are more similar to *FT* (Web site link 17). Putative orthologs of *FLC* and *AGL20*, both MADS domain genes, could not be identified among the large rice gene family. Genes from other cereals that have been shown to affect flowering time, including *Id1* in maize (*53*) and *Rht-B1/d8* in wheat and maize (*54, 55*), are represented by clear homologs in the rice genome. Although rice homologs for most of the *Arabidopsis* flowering-time genes can be identified, it is currently not clear whether the genetic network that integrates them is conserved.

Cereal and *Arabidopsis* flowers differ in perianth structure and in their arrangement on the flowering stem, but they likely develop under similar genetic control mechanisms (*56*). Rice homologs of the *Arabidopsis* floral meristem and organ identity genes *LFY*, *AP3*, *PI*, and *AG* have been described [*RFL* (*57*), *osMADS16* (*58*), *osMADS 2* and *4* (*59, 60*), and *osMADS3* (*60*), respectively]. A search of Syd with the *Arabidopsis* sequences confirmed that the rice genes previously identified are those most closely related to their *Arabidopsis* homologs. The rice *osMADS7* and *osMADS1* genes are the most similar to the *Arabidopsis* meristem identity gene *AP1* (Web site link 18). Definitive rice homologs of the *Arabidopsis CAL*, *UFO*, and *SUP* genes could not be identified within the large rice gene families. The *Arabidopsis AP2* gene has a single homolog in rice, whereas the AP2-domain gene *ANT* is represented by two homologs (Web site link 19). In agreement with previous studies demonstrating conservation of organ identity genes between monocots and dicots (*56, 60*), homologs of most *Arabidopsis* meristem and organ identity genes can be identified in rice. Functional analysis is required to demonstrate conservation of specific regulatory activities for the remaining genes (*57*).

**Metabolism.** Sequence homology suggests that about 25% of rice genes are involved in metabolism. Syd contains genes for all central metabolic processes (glycolysis; citric acid cycle; pentose phosphate pathway; photosynthesis and respiration; synthesis and degradation of amino acids, nucleotides, fatty acids, and lipids; cofactors; carbohydrates; cell wall materials) and nutrient exchange (assimilation of carbon, nitrogen, sulfur, and phosphorus; absorption of minerals). In rice, as in *Arabidopsis*, extensive gene redundancy exists across all metabolic pathways (Web site link 20). Multiple-copy genes may facilitate the tightly regulated expression of specific isoenzymes in specialized tissues, at certain developmental stages, or in response to environmental challenges (*61, 62*). Large gene families exist for enzymes putatively involved in the biosynthesis of secondary metabolites. These structurally diverse compounds are generated by only a few types of reactions (*63*), which are catalyzed by (i) enzymes forming core structures (e.g., chalcone synthase, encoded by a gene family with four members in *Arabidopsis* and 26 members in rice), (ii) redox enzymes (e.g., cytochrome P450s, encoded by >250 genes in *Arabidopsis* and >200 genes in rice), and (iii) substitution enzymes (e.g., *O*-methyltransferases, encoded by a gene family with >15 members in *Arabidopsis* and >10 members in rice). Metabolic diversity in plants is partly due to the occurrence of multifunctional enzymes. For example, terpene synthases (encoded by a gene family with >40 members in *Arabidopsis* and >15 members in rice) are known for their ability to synthesize multiple products from a single substrate (*64*), and 2-oxoglutarate–dependent dioxygenases (encoded by a gene family of >80 members in *Arabidopsis* and >50 members in rice) can typically accept multiple substrates and produce multiple products (*65*).

Specific classes of secondary metabolites produced in different plant lineages function as signal molecules, attract pollinators, or defend against herbivores and pathogens (*66*). For example, rice synthesizes sakuranetin (a flavanone); momilactones A and B (diterpenes); and oryzalexins A, B, C, D, E, F, and S (diterpenes) as the predominant antifungal phytoalexins (*67*). In contrast, pathogen-challenged *Arabidopsis* accumulates glucosinolate-derived isothiocyanates and the indole camalexin (*68, 69*). Interestingly, the genomes of rice and *Arabidopsis* contain gene families putative-

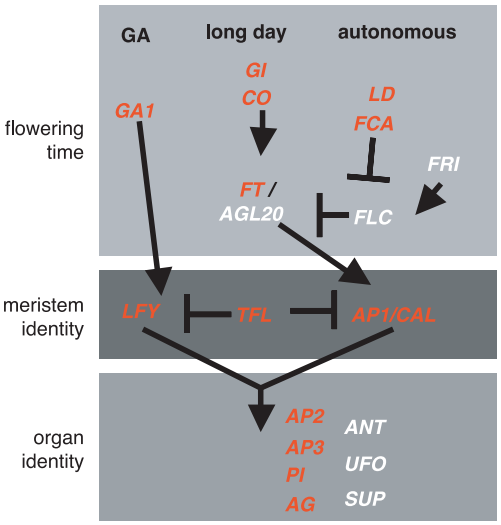



**Fig. 3.** Flowering-time and flower development genes in the rice genome. A simplified model shows the predicted genetic network regulating flowering time and flower development in *Arabidopsis*, with gene names color-coded to indicate clear identification of an ortholog in the rice genome (red) or no clear identification (white). In *Arabidopsis* there are three genetic pathways that control flowering time (*100, 101*). The long-day pathway represented by *GI* and *CO* and the autonomous pathway represented by *LD*, *FCA*, and *FLC* are likely integrated through *FT* and *AGL20* to promote activation of meristem identity genes *LFY*, *AP1*, and *CAL*. The vernalization pathway, represented by *FRI*, feeds into the autonomous pathway upstream of *FLC*. The GA pathway, represented by *GA1*, leads to the activation of *LFY*. *TFL* serves to restrict the expression of the meristem identity genes to floral meristems, where they promote the patterned expression of floral organ identity genes *AP2*, *AP3*, *PI*, and *AG*, which are also affected by the regulatory genes *ANT*, *UFO*, and *SUP* (*102, 103*).

ly encoding strictosidine β-glucosidase, berberine-bridge enzyme, and strictosidine synthase (Web site link 21). Biochemically characterized members of these enzyme families are involved in several pathways of alkaloid biosynthesis that are not known to operate in rice or *Arabidopsis* (*70*). However, repeated evolution—a process that leads to orthologous or paralogous genes with modified biochemical functions—appears to be a common theme in secondary metabolism (*66*). Hence, it is often impossible to assign the catalytic function of a novel enzyme solely on the basis of sequence similarity. Enzymes encoded by such gene families should be regarded as representatives of enzyme classes with common catalytic mechanisms (e.g., berberine-bridge enzyme is a C-C bond-forming oxidoreductase), the functions of which need to be determined biochemically.

**Phosphate transporters in the rice genome.** Improved nutrient assimilation is becoming increasingly important for modern crops. A subset of transporters involved in the acquisition and translocation of phosphate (Pi) is important for uptake of this often limiting macronutrient. Plants have both high- and low-affinity phosphate transporter systems (*71, 72*). Low-affinity phosphate transporters are constitutively expressed and operate at millimolar Pi concentrations, whereas most high-affinity phosphate transporters are transcriptionally induced at limiting phosphate concentrations and operate at micromolar concentrations. High- and low-affinity Pi transporter genes have been cloned from a number of different plant species by homology to known Pi transporters from yeast (*73, 74*). For the high-affinity transporters, six genes from *A. thaliana*, three from *Solanum tuberosum*, one from *Nicotiana tabacum*, two from *Lycopersicon esculentum*, two from *Medicago truncatula*, and one from *Catharanthus roseus* have been reported (*71, 75*). In contrast, only one low-affinity Pi transporter, preferentially expressed in leaves, has been isolated from *A. thaliana* (*76*). Plant Pi transporter genes were identified and compared by searching Syd and the *Arabidopsis* genome for homologs of *Arabidopsis* high-affinity transporter AtPT1 (*77, 78*) and low-affinity transporter Pht2;1 (*76, 79*). *Arabidopsis* and Syd were found to have 9 and 13 members, respectively, of the high-affinity Pi transporter gene family (Web site link 22). Both Syd and *Arabidopsis* have only one detectable low-affinity Pi transporter gene (Web site link 23).

**Table 5.** Transcription factor (TF) classes. Family sizes of select TF classes found in Syd are based on homology with genes in TRANSFAC (*80*).

| TF family | Number in Syd |
|---|---|
| MYB superfamily | 156 |
| Ap2/ERF/RAV | 143 |
| CH zinc finger | 125 |
| CC zinc finger | 35 |
| MADS box | 71 |
| WRKY | 83 |
| Dof | 21 |
| Trihelix | 8 |
| BZip | 75 |
| YABBY | 5 |

**Transcription factors in the rice genome.** The complement of rice transcription factors ( TFs) appears quite similar to that of *Arabidopsis* and shows many of the same overall biases of family types and numbers (*80*). A comparison ( TBLASTN) using the TRANSFAC database against *HML*genes revealed 1306 TF genes (Table 5). The number identified is similar to the 1533 TFs reported in *Arabidopsis* (*81*), although it must be an underestimate because some plant-specific TF families (e.g., Aux/IAA, NAC, GARP) are not represented in TRANSFAC. BLAST analysis was used to estimate the TF family sizes reported (*80*).

The MYB superfamily in rice is quite large relative to other sequenced organisms, with 156 readily identifiable genes, although smaller than the family of 190 MYB and MYB-related genes identified in *Arabidopsis* (*81*). The MADS box family, which also appears to have been amplified in plants, comprises 71 genes in rice, comparable to the 82 genes found in *Arabidopsis*. The C2H2 zinc finger class in rice has 125 members, close to the number in *C. elegans* and slightly more than in *Arabidopsis*.

Rice harbors all other plant-specific TF families identified in *Arabidopsis*. The C2C2 zinc finger class comprises several subtypes of proteins, including GATA/CO, Dof, and YABBY. Rice has 36 members identified as GATA/CO, whereas *Arabidopsis* has 61. The single zinc-finger Dof family comprises 21 proteins in rice and 36 in *Arabidopsis*. The YABBY family appears quite limited in rice, with only four identifiable members. Eighty-three WRKY proteins were identified in rice, slightly more than the 72 found in *Arabidopsis*. The Ap2/ERF/RAV family appears to have a similar number of members in rice (143 genes) and *Arabidopsis* (144 genes). This comparison alone provides limited functional information. However, as described above, queries with *Arabidopsis* TFs of known function can identify candidate rice orthologs with considerable similarity.

**Rice as a model for other cereals.** Extensive synteny among cereals (*7, 82–84*) allows integration of their genetic and physical maps. Sequence-based markers from syntenic regions of one cereal can be used for fine mapping and candidate gene identification across cereals. The small genome of rice provides the genomic foundation for all cereals—enabling efficient identification of orthologous genes, regulatory regions, gene functions—and may facilitate the sequencing of other cereal genomes. The extent of gene conservation was determined by compiling a set of full-length, nonredundant complete coding sequences for each nonrice cereal species, and comparing these to Syd. At significant similarity levels, almost every cereal protein was found to have a related gene in rice (Table 6). At higher stringency, 80 to 90% of cereal gene queries identified rice homologs ( Table 6). These observations suggest that most genes are conserved across cereals, and that phenotypic variation is due to a small number of different genes or functional differences within similar genes.

The level of synteny among cereals was determined by comparing anchored rice genomic sequence to mapped sequence from other cereals. Related regions of the rice and

**Table 6.** TBLASTN comparison of rice versus other cereal proteins from GenBank. A set of full-length nonredundant cereal protein sequences was compiled using all available sequences from GenBank. Pairs of proteins with greater than 90% identity over an alignment of at least 100 amino acids were considered redundant and one of the two was removed.

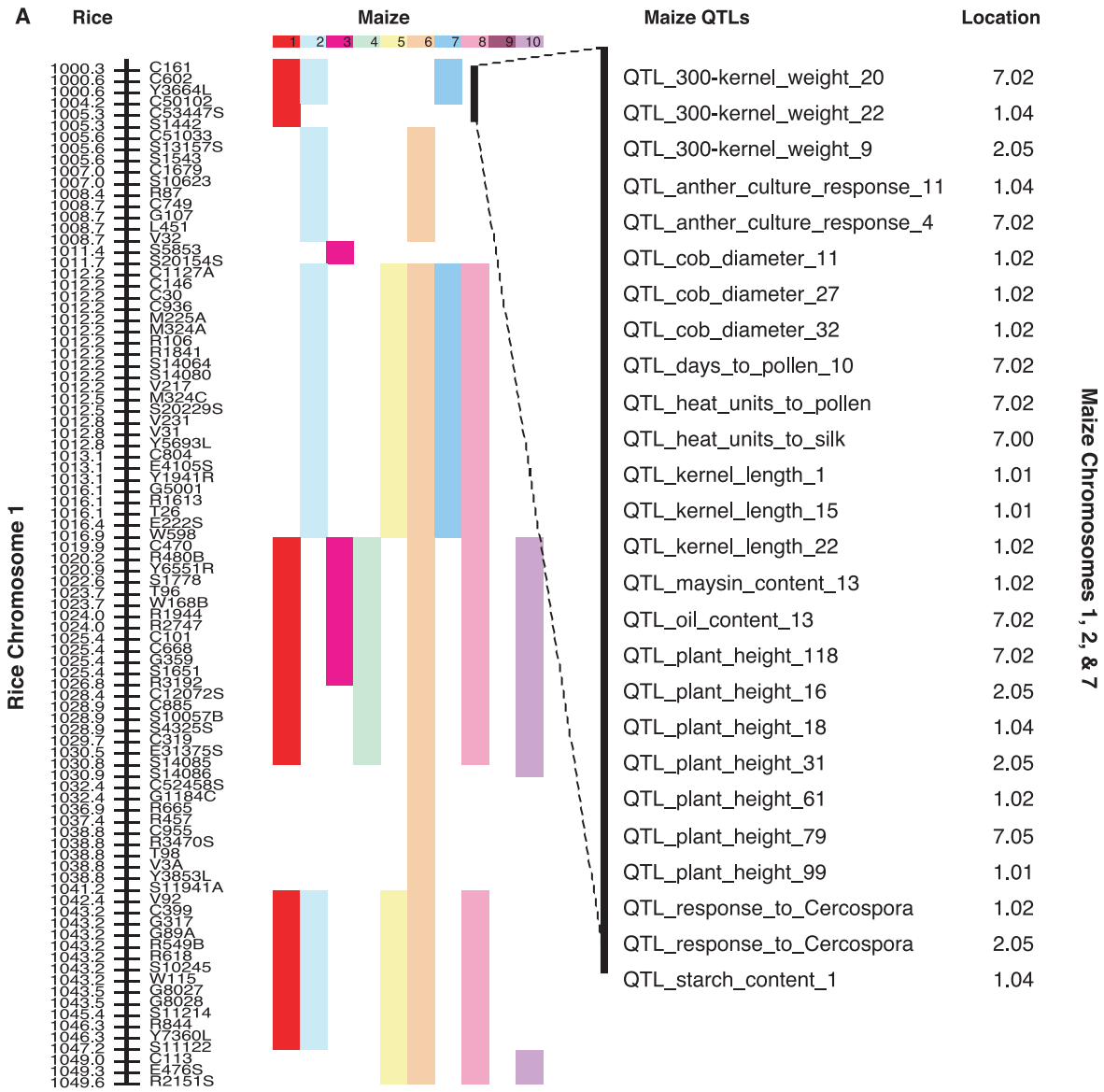
| Species | Unique coding sequence | Rice hits at | | | | | |
|---|---|---|---|---|---|---|---|
| | | $E \leq -6$ | $E \leq -9$ | $E \leq -12$ | $E \leq -25$ | $E \leq -50$ | $E \leq -100$ |
| Maize | 696 | 684 (98%) | 680 (98%) | 677 (97%) | 644 (93%) | 500 (72%) | 254 (36%) |
| Sorghum | 82 | 70 (85%) | 68 (83%) | 65 (79%) | 62 (76%) | 49 (60%) | 30 (37%) |
| Wheat | 238 | 234 (98%) | 232 (97%) | 228 (96%) | 202 (85%) | 156 (66%) | 96 (40%) |
| Barley | 296 | 293 (99%) | 293 (99%) | 292 (99%) | 273 (92%) | 220 (74%) | 116 (39%) |
| Rye | 25 | 25 (100%) | 25 (100%) | 25 (100%) | 24 (96%) | 12 (48%) | 6 (24%) |
| Oat | 44 | 42 (95%) | 42 (95%) | 42 (95%) | 37 (84%) | 35 (80%) | 20 (45%) |


**Fig. 4.** Rice-maize synteny. Maize markers were mapped to the rice genome in silico. Maize map and sequence information were derived from MaizeDB (610 markers) and GenBank, respectively. Maize chromosomes are indicated along the vertical black lines; positions of specific markers and bins are defined by horizontal lines. Rice chromosomes are represented by numbered, colored rectangles. Significant homology (at least 80% identity, over 100 continuous base pairs, between a maize chromosomal region and a particular rice region) is indicated by a colored rectangle to the right of the maize chromosome. For a more detailed version of this map, see Web site link 24.

maize genomes were aligned (Fig. 4). Significant genomic alignment was also achieved using the limited number of available sequence-based mapped markers from other cereals (Web site link 24), consistent with previous reports (*82, 85–87*). Using such alignments, traits mapped in other cereals can be associated with rice sequences facilitating identification of the underlying genes. About 2000 cereal quantitative trait loci (QTLs) have been mapped (*88–98*) and can be placed on the rice genome map en masse. For example, many maize QTLs were associated with the top of rice chromosome 1 by aligning maize chromosomes 1, 2, and 7 with this region (Fig. 5A). As a more specific example, a QTL influencing grain yield (QTL 21) that maps to maize chromosome 1 (*99*) was localized to the syntenic region of rice chromosome 3, containing ~220 *HML*genes[300] and more than 120 rice SSRs (Fig. 5B). With the use of these genes, ~100 unmapped maize cDNAs were identified by homology and are therefore candidate genes influencing yield.

**Summary.** Efficient shotgun sequencing strategies have been applied to microorganisms and recently to organisms with larger genomes such as *Drosophila*, human, and mouse. The goal of this project was to create a database of mapped cereal genes and markers, and provide a foundation for cereal functional genomics studies. The rice genome was chosen as the appropriate model cereal genome and sequenced to greater than 99% coverage and accuracy. The resulting genomic information enables development of RNA profiling, proteomics, and accelerated crop breeding technologies. Homologs of most of the known cereal

**Fig. 5.** Maize QTLs mapped to the rice genome. (**A**) Rice-maize comparative QTL mapping. Portions of maize chromosomes, represented by numbered, colored rectangles, that show sequence similarity (at least 80% identity over 100 continuous base pairs) with specific regions of the top of rice chromosome 1 are shown. The rice map is from the IRGSP. Genetic distance is indicated by the numbers to the left of the rice chromosome (e.g., 1004.2 means 4.2 cM from the tip of chromosome 1); specific markers that map to this region are indicated to the right. Regions from maize chromosomes 1, 2, and 7 show similarity with the tip of rice chromosome 1 as shown, and maize QTLs in these regions are indicated. The region represented by the thick black line comprises ~650 kbp in rice; each colored block represents varying amounts of maize DNA. (**B**) Detailed example of rice-maize comparative QTL mapping. Grain yield QTL 21 is mapped to maize map bin 1.03 between cDNA markers csu 710 and csu 392, and is syntenic with rice chromosome 3. Additional markers from the same maize bin confirm microsynteny in this target region, which contains ~220 candidate genes and 120 SSR markers in rice. Dotted lines connect homologous genes with the indicated BLAST expectation values.


A
Rice
Maize
Maize QTLs
Location
Rice Chromosome 1
Maize Chromosomes 1, 2, & 7
1 2 3 4 5 6 7 8 9 10
1000.3 C161
1000.6 C602
1000.6 Y3664L
1004.2 C50102
1005.3 C53447S
1005.3 S1442
1005.6 C51033
1005.6 S13157S
1005.6 S1543
1007.0 C1679
1007.0 S10623
1008.4 R87
1008.7 C749
1008.7 G107
1008.7 L451
1008.7 V32
1008.7 S5853
1011.4 S20154S
1011.7 C1127A
1012.2 C146
1012.2 C30
1012.2 C936
1012.2 M225A
1012.2 M324A
1012.2 R106
1012.2 R1841
1012.2 S14064
1012.2 S14080
1012.2 V217
1012.5 M324C
1012.5 S20229S
1012.8 V231
1012.8 V31
1012.8 Y5693L
1013.1 C804
1013.1 E4105S
1013.1 Y1941R
1016.1 G5001
1016.1 R1613
1016.1 T26
1016.4 E222S
1016.9 W598
1019.9 C470
1020.2 R480B
1020.9 Y6551R
1022.6 S1778
1023.7 T96
1023.7 W168B
1024.0 R1944
1024.0 R2747
1025.4 C101
1025.4 C668
1025.4 G359
1025.4 S1651
1026.8 R3192
1028.4 C12072S
1028.9 C885
1028.9 S10057B
1028.9 S4325S
1029.7 C319
1030.5 E31375S
1030.8 S14085
1030.9 S14086
1032.4 C52458S
1032.4 G1184C
1036.9 R665
1037.4 R457
1038.8 C955
1038.8 R3470S
1038.8 T98
1038.8 V3A
1038.8 Y3853L
1041.2 S11947A
1042.4 V92
1043.2 C399
1043.2 G317
1043.2 G89A
1043.2 R549B
1043.2 R618
1043.2 S10245
1043.2 W115
1043.5 G8027
1043.5 G8028
1045.4 S11214
1046.3 R844
1046.3 Y7360L
1047.2 S11122
1049.0 C113
1049.3 E476S
1049.6 R2151S
QTL_300-kernel_weight_20 7.02
QTL_300-kernel_weight_22 1.04
QTL_300-kernel_weight_9 2.05
QTL_anther_culture_response_11 1.04
QTL_anther_culture_response_4 7.02
QTL_cob_diameter_11 1.02
QTL_cob_diameter_27 1.02
QTL_cob_diameter_32 1.02
QTL_days_to_pollen_10 7.02
QTL_heat_units_to_pollen 7.02
QTL_heat_units_to_silk 7.00
QTL_kernel_length_1 1.01
QTL_kernel_length_15 1.01
QTL_kernel_length_22 1.02
QTL_maysin_content_13 1.02
QTL_oil_content_13 7.02
QTL_plant_height_118 7.02
QTL_plant_height_16 2.05
QTL_plant_height_18 1.04
QTL_plant_height_31 2.05
QTL_plant_height_61 1.02
QTL_plant_height_79 7.05
QTL_plant_height_99 1.01
QTL_response_to_Cercospora 1.02
QTL_response_to_Cercospora 2.05
QTL_starch_content_1 1.04
B
Maize QTL – Grain Yield 21
24cM
umc1403
umc1397
umc1479
QTL
Maize Bin 1.03
Maize IBM Map
Chrom 1
14cM
csu179a
ccsu238a
csu710
csu392
Maize UMC 98 Map
Chrom 1
4.0e-33
1e-130
3e-53
4e-22
2e-30
2e-19
2e-45
Rice Chrom. 3
37cM
41cM
44cM
45cM
46cM
47cM
49cM
~ 1.3 MB
220 Predicted Genes
120 Trinucleotide SSRs
~ 4 MB
734 Predicted Genes

proteins were found in rice. Homologs of most predicted *Arabidopsis* proteins were also identified, although synteny between rice and *Arabidopsis* is limited. Several thousand genes were found to be present only in the *Arabidopsis* and rice genomes, and are candidates for plant-specific genes. Many rice genes were assigned putative roles via comparison with *Arabidopsis* genes. Biosynthetic enzymes, signal transduction proteins, developmental regulators, and specific ion transporters were readily identified in the rice genome. Assembled sequence data were aligned to the rice physical and genetic maps, and anchored to heterologous cereal maps. The resulting universal cereal map allows placement of most mapped cereal QTLs and assignment of trait candidate genes. The draft genome sequence described here provides a foundation for cereal genomics; however, highly accurate, finished sequence should remain the ultimate goal for plant science. Continued application of genomic and biotechnology tools to crop improvement will be necessary to meet future food, health, and material challenges.

### References and Notes


1. J. R. Harlan, *The Living Fields: Our Agricultural Heritage* (Cambridge Univ. Press, New York, 1995), pp. 30–31.
2. World Agricultural Supply and Demand Estimates (WASDE), http://usda.mannlib.cornell.edu/reports/waobr/wasde-bb/2001/wasde377.txt.
3. G. S. Khush, *Plant Mol. Biol.* **35**, 25 (1997).
4. E. A. Kellogg, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 2005 (1998).
5. M. D. Gale, K. M. Devos, *Science* **282**, 656 (1998).
6. V. L. Chandler, S. Wessler, *Plant Physiol.* **125**, 1155 (2001).
7. M. Freeling, *Plant Physiol.* **125**, 1191 (2001).
8. T. Sasaki, B. Burr, *Curr. Opin. Plant Biol.* **3**, 138 (2000).
9. National Center for Biotechnology Information, Database of Expressed Sequence Tags (www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html).
10. G. F. Barry, *Plant Physiol.* **125**, 1164 (2001).
11. J. Yu, S. Hu, J. Wang, J., S. Li, *Chin. Sci. Bull.* **46**, 1937 (2001).
12. K. H. Wolfe, M. Gouy, Y. W. Yang, P. M. Sharp, W. H. Li, *Proc. Natl. Acad. Sci. U.S.A.* **86**, 6201 (1989).
13. The *Arabidopsis* Genome Initiative, *Nature* **408**, 796 (2000).
14. H. M. Goodman, J. R. Ecker, C. Dean, *Proc. Natl. Acad. Sci. U.S.A.* **92**, 10831 (1995).
15. R. J. Cook, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 1993 (1998).
16. M. D. Adams *et al.*, *Science* **287**, 2185 (2000).
17. J. C. Venter *et al.*, *Science* **291**, 1304 (2001).
18. J. C. Venter *et al.*, *Science* **280**, 1540 (1998).
19. G. G. Presting *et al.*, *Novartis Found. Symp.* **236**, 13 (2001).
20. L. Mao *et al.*, *Genome Res.* **10**, 982 (2000).
21. M. Chen *et al.*, *Plant Cell* **14**, 1 (2002).
22. R. A. Wing *et al.*, in *Rice Genetics IV, Proceedings of the Fourth International Rice Genetics Symposium,* G. S. Khush, D. S. Brar, B. Hardy, Eds. (IRRI Press, Makati City, Philippines, 2001), pp. 215–225.
23. About 80% of the sequences were from paired (forward and reverse) reads with an average clone size of ~1700 bp (18.5-fold genome coverage). More than fivefold coverage was from randomly selected clones, with the remainder from resequencing gaps or low-quality regions. Low-voltage electrophoresis was used for resequencing, which provided longer sequences with better quality and in many cases resulted in closing gaps between contigs. The resulting sequences were analyzed for contamination from nonrice DNA sources (~500,000 reads) or rice repetitive DNA (~1,500,000 reads) and the remainder assembled using the Myriad Assembly Program.
24. J. B. Hogenesch *et al.*, *Cell* **106**, 413 (2001).
25. K. Hofmann, P. Bucher, L. Falquet, A. Bairoch, *Nucleic Acids Res.* **27**, 215 (1999).
26. A. Bateman *et al.*, *Nucleic Acids Res.* **28**, 263 (2000).
27. R. Apweiler *et al.*, *Nucleic Acids Res.* **29**, 37 (2001).
28. R. Apweiler *et al.*, *Bioinformatics* **16**, 1145 (2000).
29. S. Lewis, M. Ashburner, M. G. Reese, *Curr. Opin. Struct. Biol.* **10**, 349 (2000).
30. The *C. elegans* Sequencing Consortium, *Science* **282**, 2012 (1998).
31. Y. Harushima *et al.*, *Genetics* **148**, 479 (1998).
32. S. J. O'Brien *et al.*, *Nature Genet.* **3**, 103 (1993).
33. T. H. Lan *et al.*, *Genome Res.* **10**, 776 (2000).
34. Rice genome duplications were dated by calculating amino acid divergence rates of all possible paralogous protein pairs. 14,345 high-evidence rice proteins were grouped by chromosomes. Paralogous protein pairs were identified by comparing groups (BLASTP). Protein pairs are defined as those with 80% identity over a minimum of 30 amino acids. Protein pairs were aligned with CLUSTALW, and amino acid divergence rates ($d_A$) were estimated by PAML (Phylogenetic Analysis by Maximum Likelihood, version 3.0, University College, London) using the Dayhoff matrix. The divergence time calculation was based on a molecular clock rate of $9 \times 10^{-10}$ nonsynonymous substitutions per site per lineage per year and 2.25 nonsynonymous substitutions per amino acid change.
35. T. J. Vision, D. G. Brown, S. D. Tanksley, *Science* **290**, 2114 (2000).
36. M. O. Dayhoff, R. M. Schwartz, B. C. Orcutt, *Atlas of Protein Sequence and Structure, Vol. 5* (National Biomedical Research Foundation, Washington, DC, 1978), pp. 345–352.
37. S. V. Muse, *Plant Mol. Biol.* **42**, 25 (2000).
38. B. S. Gaut, J. F. Doebley, *Proc. Natl. Acad. Sci. U.S.A.* **94**, 6809 (1997).
39. W. A. Wilson *et al.*, *Genetics* **153**, 453 (1999).
40. S. Mathews, R. A. Sharrock, *Mol. Biol. Evol.* **13**, 1141 (1996).
41. A. M. van Dodeweerd *et al.*, *Genome* **42**, 887 (1999).
42. K. Mayer *et al.*, *Genome Res.* **11**, 1167 (2001).
43. A. H. Paterson *et al.*, *Nature Genet.* **14**, 380 (1996).
44. *Arabidopsis* annotated proteins of chromosomes 1, 2, and 4 were obtained from GenBank, and annotated proteins of chromosomes 3 and 5 were obtained from The Institute for Genomic Research (TIGR) (May 2001). *Arabidopsis* proteins from each chromosome were compared to anchored rice sequence contigs by BLAST, effectively linking the *Arabidopsis* and rice maps and enabling a study of syntenic relationships between the two species. Requiring at least 70% identity over a minimum of 30 contiguous amino acids, 98% of BLAST hits achieved E values of $\leq -7$. Syntenic groups are defined as three or more *Arabidopsis* proteins from the same chromosome mapping to one rice BAC contig. Bootstrap analysis was used to determine the significance threshold (Table 4).
45. H. M. Ku, T. Vision, J. Liu, S. D. Tanksley, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 9121 (2000).
46. D. Grant, P. Cregan, R. C. Shoemaker, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 4168 (2000).
47. J. L. Dangl, J. D. Jones, *Nature* **411**, 826 (2001).
48. BLAST E score $< -3$, searching the draft sequence with the pfam0093 NB-ARC consensus sequence as the query.
49. J. Glazebrook, *Curr. Opin. Plant Biol.* **4**, 301 (2001).
50. M. Yano *et al.*, *Plant Cell* **12**, 2473 (2000).
51. L. Pnueli *et al.*, *Development* **125**, 1979 (1998).
52. D. Bradley *et al.*, *Nature* **379**, 791 (1996).
53. J. Colasanti, Z. Yuan, V. Sundaresan, *Cell* **93**, 593 (1998).
54. J. Peng *et al.*, *Nature* **400**, 256 (1999).
55. J. M. Thornsberry *et al.*, *Nature Genet.* **28**, 286 (2001).
56. B. A. Ambrose *et al.*, *Mol. Cell* **5**, 569 (2000).
57. J. Kyozuka, S. Konishi, K. Nemoto, T. Izawa, K. Shimamoto, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 1979 (1998).
58. Y. H. Moon, J.-Y. Jung, H. G. Kang, G. An, *Plant Mol. Biol.* **40**, 167 (1999).
59. Y. Y. Chung *et al.*, *Plant Sci.* **109**, 45 (1995).
60. H. G. Kang, J. S. Jeon, S. Lee, G. An, *Plant Mol. Biol.* **38**, 1021 (1998).
61. B. M. Lange, T. Rujan, W. Martin, R. Croteau, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 13172 (2000).
62. R. A. Dixon, *Nature* **411**, 843 (2001).
63. P. J. Facchini, *Trends Plant Sci.* **4**, 382 (1999).
64. J. Bohlmann, G. Meyer-Gauen, R. Croteau, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 4126 (1998).
65. A. G. Prescott, P. John, *Annu. Rev. Plant Physiol.* **47**, 245 (1996).
66. E. Pichersky, D. R. Gang, *Trends Plant Sci.* **5**, 439 (2000).
67. R. J. Grayer, T. Kokubun, *Phytochemistry* **56**, 253 (2001).
68. K. F. Tierens *et al.*, *Plant Physiol.* **125**, 1688 (2001).
69. J. Tsuji, E. Jackson, D. Gage, R. Hammerschmidt, S. Somerville, *Plant Physiol.* **98**, 1304 (1992).
70. *Dictionary of Natural Products on CD-ROM* (Chapman & Hall/CRC Press, Boca Raton, FL, 2000).
71. K. G. Raghothama, *Trends Plant Sci.* **5**, 412 (2000).
72. M. J. Chrispeels, N. M. Crawford, J. I. Schroeder, *Plant Cell* **11**, 661 (1999).
73. M. Bun-Ya, M. Nishimura, S. Harashima, Y. Oshima, *Mol. Cell. Biol.* **11**, 3229 (1991).
74. P. Martinez, R. Zvyagilskaya, P. Allard, B. L. Persson, *J. Bacteriol.* **180**, 2253 (1998).
75. C. Rausch *et al.*, *Nature* **414**, 462 (2001).
76. P. Daram *et al.*, *Plant Cell* **11**, 2153 (1999).
77. GenBank accession number U62330.
78. U. S. Muchhal, J. M. Pardo, K. G. Raghothama, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 10519 (1996).
79. GenBank accession number X98130.
80. The 3501 TFs in the TRANSFAC data set (v5.2) were compared against the rice gene predictions (no size cutoff) using TBLASTN. Only matches with an E value $\leq -4$ and in which the subject extended at least 70% of the length of the TF-specific motif or domain in the query were included. In a parallel analyses of the *Arabidopsis* genome, 1799 TF genes were identified.
81. J. L. Riechmann *et al.*, *Science* **290**, 2105 (2000).
82. M. D. Gale, K. M. Devos, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 1971 (1998).
83. M. Gale, G. Moore, K. Devos, *Novartis Found. Symp.* **236**, 46 (2001).
84. C. Feuillet, B. Keller, *Proc. Natl. Acad. Sci. U.S.A.* **96**, 8265 (1999).
85. J. L. Bennetzen, M. Freeling, *Genome Res.* **7**, 301 (1997).
86. S. R. McCouch, *Plant Physiol.* **125**, 152 (2001).
87. B. Keller, C. Feuillet, *Trends Plant Sci.* **5**, 246 (2000).
88. I. J. Havukkala, *Curr. Opin. Genet. Dev.* **6**, 711 (1996).
89. M. Lee, *Symp. Soc. Exp. Biol.* **50**, 31 (1996).
90. S. R. McCouch, R. W. Doerge, *Trends Genet.* **11**, 482 (1995).
91. J. Doebley, A. Stec, C. Gustus, *Genetics* **141**, 333 (1995).
92. J. Xiao, J. Li, L. Yuan, S. D. Tanksley, *Genetics* **140**, 745 (1995).
93. J. C. Lanceras *et al.*, *DNA Res.* **7**, 93 (2000).
94. T. J. Flowers *et al.*, *J. Exp. Bot.* **51**, 99 (2000).
95. M. Yano, T. Sasaki, *Plant Mol. Biol.* **35**, 145 (1997).
96. A. H. Paterson, *Genome Res.* **5**, 321 (1995).
97. A. E. Melchinger, H. F. Utz, C. C. Schon, *Genetics* **149**, 383 (1998).
98. S. R. McCouch *et al.*, *Plant Mol. Biol.* **35**, 89 (1997).
99. L. R. Veldboom, M. Lee, *Crop Sci.* **36**, 1310 (1996).
100. P. H. Reeves, G. Coupland, *Curr. Opin. Plant Biol.* **3**, 37 (2000).
101. T. Araki, *Curr. Opin. Plant Biol.* **4**, 63 (2001).
102. O. J. Ratcliffe *et al.*, *Development* **125**, 1609 (1998).
103. D. Weigel, E. M. Meyerowitz, *Cell* **78**, 203 (1994).


104. We thank D. Patton, J. Salmeron, B. Dietrich, A. Binder, and L. Mattle for critical reading of the manuscript, and S. Guimil for artwork.

21 November 2001; accepted 7 March 2002

# ERRATUM

post date 5 August 2005

**Research Articles:** "A draft sequence of the rice genome (*Oryza sativa* L. spp. *japonica*)" by S. A. Goff *et al*. (5 Apr. 2002, p. 92). Syngenta's rice genome sequence information was donated to the International Rice Genome Sequencing Project (IRGSP) and the Beijing Genomics & Bioinformatics Institute (BGI) in 2002 under an agreement to integrate the Syngenta data with the sequence generated by these projects and deposit the assemblies into appropriate public databases, such as GenBank and EMBL. The individual sequence reads from Syngenta were deposited into the GenBank Trace Database by the BGI in February 2004. The Syngenta draft sequences can be obtained by request at the Web site www.tmri.org/en/partnership/access.aspx, although the authors recommend the use of the more complete sequence now in GenBank.