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

Jun Yu,[1,2,3,4*] Songnian Hu,[1*] Jun Wang,[1,2,5*] Gane Ka-Shu Wong,[1,2,4*] Songgang Li,[1,5] Bin Liu,[1] Yajun Deng,[1,6] Li Dai,[1] Yan Zhou,[2,7] Xiuqing Zhang,[1,3] Mengliang Cao,[8] Jing Liu,[2] Jiandong Sun,[1] Jiabin Tang,[1,3] Yanjiong Chen,[1,6] Xiaobing Huang,[1] Wei Lin,[2] Chen Ye,[1] Wei Tong,[1] Lijuan Cong,[1] Jianing Geng,[1] Yujun Han,[1] Lin Li,[1] Wei Li,[1,9] Guangqiang Hu,[1] Xiangang Huang,[1] Wenjie Li,[1] Jian Li,[1] Zhanwei Liu,[1] Long Li,[1] Jianping Liu,[1] Qiuhui Qi,[1] Jinsong Liu,[1] Li Li,[1] Tao Li,[1] Xuegang Wang,[1] Hong Lu,[1] Tingting Wu,[1] Miao Zhu,[1] Peixiang Ni,[1] Hua Han,[1] Wei Dong,[1,3] Xiaoyu Ren,[1] Xiaoli Feng,[1,3] Peng Cui,[1] Xianran Li,[1] Hao Wang,[1] Xin Xu,[1] Wenxue Zhai,[3] Zhao Xu,[1] Jinsong Zhang,[3] Sijie He,[3] Jianguo Zhang,[1] Jichen Xu,[3] Kunlin Zhang,[1,5] Xianwu Zheng,[3] Jianhai Dong,[2] Wanyong Zeng,[3] Lin Tao,[2] Jia Ye,[2] Jun Tan,[2] Xide Ren,[1] Xuewei Chen,[3] Jun He,[2] Daofeng Liu,[3] Wei Tian,[2,6] Chaoguang Tian,[1] Hongai Xia,[1] Qiyu Bao,[1] Gang Li,[1] Hui Gao,[1] Ting Cao,[1] Juan Wang,[1] Wenming Zhao,[1] Ping Li,[3] Wei Chen,[1] Xudong Wang,[3] Yong Zhang,[1,5] Jianfei Hu,[1,5] Jing Wang,[1,5] Song Liu,[1] Jian Yang,[1] Guangyu Zhang,[1] Yuqing Xiong,[1] Zhijie Li,[1] Long Mao,[3] Chengshu Zhou,[8] Zhen Zhu,[3] Runsheng Chen,[1,9] Bailin Hao,[2,10] Weimou Zheng,[1,10] Shouyi Chen,[3] Wei Guo,[11] Guojie Li,[12] Siqi Liu,[1,2] Ming Tao,[1,2] Jian Wang,[1,2] Lihuang Zhu,[3†] Longping Yuan,[8†] Huanming Yang[1,2,3†]

We have produced a draft sequence of the rice genome for the most widely cultivated subspecies in China, *Oryza sativa* L. ssp. *indica*, by whole-genome shotgun sequencing. The genome was 466 megabases in size, with an estimated 46,022 to 55,615 genes. Functional coverage in the assembled sequences was 92.0%. About 42.2% of the genome was in exact 20-nucleotide oligomer repeats, and most of the transposons were in the intergenic regions between genes. Although 80.6% of predicted *Arabidopsis thaliana* genes had a homolog in rice, only 49.4% of predicted rice genes had a homolog in *A. thaliana*. The large proportion of rice genes with no recognizable homologs is due to a gradient in the GC content of rice coding sequences.

Rice is the most important crop for human consumption, providing staple food for more than half the world's population. The euchromatic portion of the rice genome is estimated to be 430 Mb in size (*1–3*), which is the smallest of the cereal crops. It is 3.7 times larger than that of *A. thaliana* (*4–6*), and 6.7 times smaller than that of the human (*7*, *8*). The well-established protocols for high-efficiency genetic transformation, widespread availability of high-density genetic and physical maps (*9*, *10*), and high degrees of synteny among cereal genomes (*11–15*) combine to make rice a unique organism for studying the physiology, developmental biology, genetics, and evolution of plants. The International Rice Genome Sequencing Project (IRGSP) (*16*) has already delivered a substantial amount of sequence for the *japonica* (*Nipponbare*) subspecies, in bacterial artificial chromosome (BAC) and P1-derived artificial chromosome (PAC)–sized contigs. Working independently, Monsanto and Syngenta (*17*, *18*) established proprietary working drafts for *japonica*, in April 2000 and February 2001, respectively. The Monsanto sequence has been used to assist in the efforts of the IRGSP.

We are releasing a draft genome sequence for rice from *93-11* (*19*), which is a cultivar of *Oryza sativa* L. ssp. *indica*, the major rice subspecies grown in China and many other Asia-Pacific regions. It is the paternal cultivar of a super-hybrid rice, *Liang-You-Pei-Jiu* (*LYP9*), which has 20 to 30% more yield per hectare than the other rice crops in cultivation (*20*). The maternal cultivar of *LYP9* is *Pei-Ai 64s* (*PA64s*), which has a major background of *indica* and a minor background of *japonica* and *javanica*, two other commonly cultivated subspecies. We have also produced a low-coverage draft sequence for *PA64s*. A preliminary assembly and analysis on a subset of this sequence was published in the *Chinese Science Bulletin* (*21*). Our discussion will focus largely on the genome landscape of rice, how it differs from that of the other sequenced plant, *A. thaliana*, and how both plant genomes differ from that of the human. We will show that rice genes exhibit a gradient in GC content, codon usage, and amino acid usage. This compositional gradient reflects a unique phenomenon in the evolutionary history of rice, and perhaps all monocot plants, but not eudicot plants. As a result, about one-half of the predicted rice genes have no obvious homolog in *A. thaliana*, whereas the other half is almost a replica of the *A. thaliana* gene set.

The entire rice genome sequence can be downloaded from our Web site at http://btn.genomics.org.cn/rice. Following our announcement of the rice genome sequence at the annual Plant, Animal and Microbe Genomes (PAG X) conference, in San Diego, during the ensuing period from 14 January to 2 March 2002, this sequence was downloaded 556 times, and the BLAST search facilities were used 7008 times by 343 individuals. This sequence has also been deposited at the DNA Data Bank of Japan/European Molecular Biology Laboratory/GenBank under the project accession number AAAA00000000. The version described in this paper is AAAA01000000.

Experimental design. The rice genome project at the Beijing Genomics Institute has been designed in two stages. This is a report on stage I, the primary objective of which was to generate a draft sequence of rice at ~4× coverage for *93-11*. A similar amount of data will eventually be generated for

[1]Beijing Genomics Institute/Center of Genomics and Bioinformatics, Chinese Academy of Sciences, Beijing 101300, China. [2]Hangzhou Genomics Institute–Institute of Bioinformatics of Zhejiang University–Key Laboratory of Bioinformatics of Zhejiang Province, Hangzhou 310007, China. [3]Institute of Genetics, Chinese Academy of Sciences, Beijing 100101, China. [4]University of Washington Genome Center, Department of Medicine, Seattle, WA 98195, USA. [5]College of Life Sciences, Peking University, Beijing 100871, China. [6]Medical College, Xi'an Jiaotong University, Xi'an 710061, China. [7]Fudan University, Shanghai 200433, China. [8]National Hybrid Rice R&D Center, Changsha 410125, China. [9]Laboratory of Bioinformatics, Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101, China. [10]Institute of Theoretical Physics, Chinese Academy of Sciences, Beijing 100080, China. [11]Digital China Ltd., Beijing 100080, China. [12]Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, China.

*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: hyang@genomics.org.cn.

PA64s, but at present there is only enough data to estimate polymorphism rates between rice cultivars. The sequence reads were acquired on high-throughput capillary machines (MegaBACE 1000, 10 to 11 runs per machine per day). Concurrent with the data acquisition, we developed a software package (*22*) to identify and mask repetitive sequences and to correctly assemble these sequence reads into contigs and scaffolds, even though cereal genomes contain far more repetitive sequence than many other genomes (*23*, *24*). We generated 87,842 expressed sequence tags (ESTs), against our ultimate goal of 1,000,000 ESTs, to provide confirmatory evidence for gene identification, and for gene expression analysis. Comparing the *93-11* contig assemblies with the public data, we generated a set of polymorphic markers for genetic analysis. In stage II of the project, our objective will be to obtain a high-quality sequence, fully integrated with the physical/genetic maps, and with complete gene annotations.

We used a "whole-genome shotgun" approach, as successfully applied to *Drosophila melanogaster* (*25*) and *Homo sapiens* (*8*). Our data are complementary to those of the IRGSP, which is sequencing *Nipponbare*, a cultivar of the subspecies *japonica*, with a "clone-by-clone" approach. If we assume a euchromatic rice genome size of 430 Mb, and a Phred Q20 (*26*, *27*) read length of 500 base pairs (bp), then 1× coverage would be equivalent to 0.86 million sequence reads, or 1 million reads after the typical success rate of 80 to 85% is factored in. Shotgun libraries were constructed with a variety of methods for clone-insert preparation (*28*–*30*), to minimize the likelihood of systematic biases in genome representation. A total of 55 plasmid libraries were constructed for *93-11* and *PA64s*, with a 2-kb nominal clone-insert size. Overall, we prepared 2.75 million plasmid DNA samples (*31*, *32*). Sequencing was performed on both ends of the inserts. By the 21 October 2001 freeze, there were 4.62 million successful reads, indicating an 84% success rate. The average Q20 read length was 546 bp.

Assembling the draft. Genomic studies of grasses, especially the cereal crops, have indicated that the intergenic regions between genes are inhabited by clusters of nested retrotransposons (*23*, *24*, *33*), which compose almost half of the rice genome, and substantially larger fractions of other crop plants like *Zea mays* (maize) and *Triticum aestivum* (wheat). Our sequence assembler software was designed to handle highly repetitive genomes without having to first characterize the repeats in any traditional biological sense. The focus was on contiguity at the scaffold level, instead of complete assembly across all the repeats. However, error probabilities would be computed for every base that was successfully assembled.

A typical assembly, based on our software RePS (Repeat-masked Phrap with Scaffolding) (*22*), is shown in Fig. 1. We began by computing the number of times that any 20-bp sequence (20-nucleotide oligomer, 20-mer) appeared in the data set. Those 20-mers that appeared more often than a fixed threshold were flagged as mathematically defined repeats (MDRs). RePS made no effort to identify biologically defined repeats (BDRs), because if a 20-mer was repeated in the MDR sense, it would complicate the sequence assembly, regardless of its biological context (e.g., microsatellites, transposable elements, multigene families, recently duplicated chromosomal segments, or pseudogenes). Instead, it masked the MDRs, so that they were invisible to the sequence assembler Phrap (*34*). This reduced the computational load by many orders of magnitude, while minimizing the likelihood of making a false join. However, it also introduced another class of gaps, repeat masked gaps (RMGs), distinct from the Lander-Waterman gaps (LWGs) that are usually encountered in sequencing. In a RMG, the gap sequence is actually in the data set, but it was not usable because it was made invisible to Phrap by the masking. In a LWG, the gap sequence is missing, as a result of sampling statistics (*35*). Some of the RMGs could be closed with the clone-end pairing information, assuming that both clone ends were not fully masked. After repeat-gap closure, and regardless of the nature of the remaining gaps, RePS was used to analyze the clone-end pairing information to construct scaffolds—nonoverlapping contigs linked together in the correct order and orientation. LWGs were usually small, easy to close by polymerase chain reaction. Gaps larger than a few kb were usually RMGs due to the nested retrotransposons in the

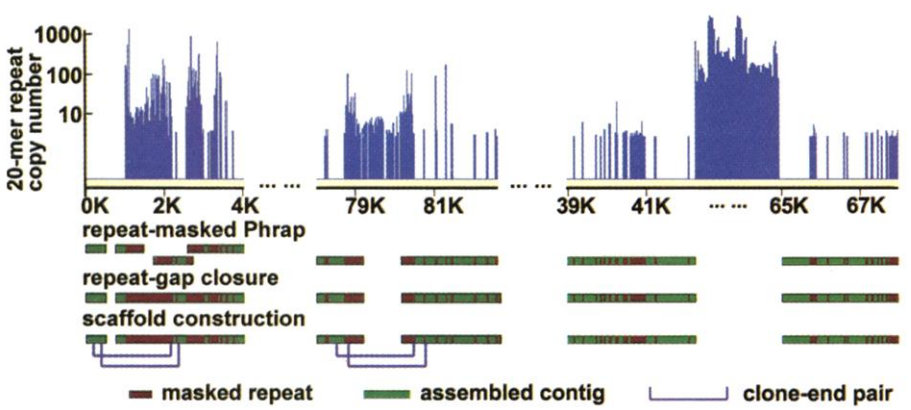

Fig. 1. Typical RePS assembly, with *93-11* (*indica*) contigs aligned to finished BAC sequences from *GLA* (*indica*) (GenBank accession numbers AL442007 and AL512542). Exact 20-mer repeats are indicated by the blue histogram bars, with bar heights proportional to estimated copy number in *93-11* (*indica*). Three stages are shown: repeat-masked Phrap, repeat-gap closure, and scaffold construction. First, we mask exact the 20-mer repeats and use Phrap to assemble the data on the basis of the unique sequence. Second, we use the clone-end pairing information to close smaller repeat masked gaps (RMGs) ignored by Phrap because of the masking. However, larger RMGs and gaps due to sampling statistics, Lander-Waterman gaps (LWGs), cannot be so closed. Third, we use the clone-end pairing information to construct scaffolds—sets of nonoverlapping contigs linked together in the correct order and orientation. A LWG at 0.5 kb is scaffolded over. RMGs at 1.5 and 2.5 kb are closed, and another at 80 kb is scaffolded over. The RMG between 42 and 65 kb is too large to scaffold across given a clone-insert size of 2 kb.

Table 1. Sequence assembly statistics for *93-11* (*indica*). The Q20 read lengths refer to the usable part of the sequence with error probabilities less than 10^{-2}. Masking 20-mer repeats eliminated 42.2% of the sequence by length. Some reads were partially masked, but 18.7% of reads were fully masked. The N50 contig or scaffold sizes define that size above which 50% of the assembly was found. To estimate the assembled-equivalent size of the unused reads, we divided total Q20 lengths by the 4.2× depth of reads in the assembled contigs. This resulted in an assembled-equivalent size of 104 Mb, of which 78 Mb was fully masked reads. The total genome size was thus estimated to be 466 Mb.

Basic shotgun data	
Total genome size (Mb)	466
Number of reads	3,565,386
Q20 read lengths (bp)	546
Shotgun coverage	4.2
Exact 20-nt oligomer repeats	
Length of fraction masked	42.2%
No. of fully masked reads	18.7%
Sequence assembly	
Total contig size (Mb)	361
N50 contig size (kb)	6.69
Total scaffold size (Mb)	362
N50 scaffold size (kb)	11.76
Unassembled data	
Fully masked reads (Mb)	78
All other reads (Mb)	26

intergenic regions between genes. Whether these gaps should be closed or not remains to be resolved.

Shotgun data for *93-11* (Table 1) and *PA64s* were assembled separately, to allow for large differences in their genome sequences. In *93-11*, there were 3.57 million sequence reads after removal of the ones containing mitochondrial, chloroplastic, and bacterial sequence. Our RePS assembly yielded 127,550 contigs with an N50 size (i.e., the size above which 50% of the total assembly is found) of 6.69 kb. The total contig length was 361 Mb. These contigs were linked into 103,044 scaffolds with an N50 size of 11.76 kb, or a 1.8-fold increase over the initial contigs. The total scaffold length was 362 Mb. In contrast, for the *PA64s* data set, we had only 1.05 million sequence reads. With such low coverage, the N50 contig and scaffold sizes were much smaller, at 1.88 and 1.97 kb, respectively. These statistics differ slightly from those reported in the *Chinese Science Bulletin* (*21*), because of improvements in the RePS software. Remaining gaps between scaffolds are probably larger than the clone-insert size of 2 kb; otherwise, we would have been able to bridge them. We cannot provide a gap size distribution, but in the rice BACs that have been sequenced, repeat cluster sizes up to 25 kb have been observed.

The total contig and scaffold lengths fall far short of the previously estimated euchromatic genome size of 430 Mb. Where is the missing DNA? In the initial phase of the RePS assembly, 42.2% of the sequence was identified as a MDR and masked. A total of 18.7% of all the reads were fully masked and not immediately usable. Even though some were later incorporated into the assembly, with the clone-end pairing information, a large number of fully masked reads, and some partially masked reads, remained unused. To estimate the effective-assembled size of the unused reads, we defined an empirical coverage based on the depth of reads in the assembled contigs, 4.2×. The effective-assembled size for the unused fully masked and partially masked reads was thus estimated as 78 and 26 Mb, respectively, resulting in a total genome size of 466 Mb. That this is larger than the previous estimates is reasonable, given that whole-genome shotgun data inevitably contain some amount of heterochromatin DNA.

Quality assessments. We assumed that any large cluster of MDRs was an intergenic region and that we could safely avoid having to assemble across such a region. If so, then most of the "functional sequence" that encodes genes, and their immediate regulatory elements, should lie in our 361 Mb of assembled contigs. To confirm that this was indeed the case, we gathered all the publicly available sequence-tagged sites (STSs) and full-length cDNA sequences, as well as our own ESTs, and searched for them in our assembled contigs, using BLAST (*36*). Fortunately, a dense physical map of STS markers had already been established (*37*) for *japonica*. A total of 2845 markers were analyzed, and on the basis of sequence identity, 91.5% of their total length could be found in our contigs. Similarly, 24,776 UniGene clusters were assembled from 87,842 ESTs for *93-11*, and 93.8% of their total length could be found in our contigs. Finally, 907 nonredundant cDNA sequences were extracted from GenBank release 125 (15 August 2001), and 90.8% of their total length could be found in our contigs. Averaged across these three data sets, the functional coverage was 92.0%.

The quality metrics that matter for gene identification are (i) contiguity on the length scale of a gene, (ii) single-base error probability, and (iii) contig assembly accuracy on the length scale of a gene. As will be detailed in a later section, the mean gene size for rice is about 4.5 kb. Considering that our N50 scaffold size is only 11.76 kb, larger scaffolds would reduce the number of genes that are split across scaffolds, and this is a key objective in stage II of the project. The number most often cited is the single-base error probability, which the International Human Genome Sequencing Consortium (*7*) determined should be 10^{-4} or better, based on a human polymorphisms rate of 10^{-3}. Actually, as is detailed in a later section, rice polymorphism rates are closer to 10^{-2}, so an error rate of 10^{-4} is better than needed. On the basis of Phrap estimates (*26*, *27*, *34*), 94.2, 90.8, and 83.5% of the *93-11* sequence had an error rate of better than 10^{-2}, 10^{-3}, and 10^{-4}, respectively. However, most of the problematic bases were at the ends of the contigs. When we restricted this calculation to contigs greater than 3 kb and ignored bases within 500 bp of the ends, 97.3, 96.1, and 92.5% of the *93-11* sequence had an error rate of better than 10^{-2}, 10^{-3}, and 10^{-4}, respectively. It is important to bear these error rates in mind when comparing two sequences to estimate polymorphism rates.

Assembly accuracy is an often overlooked but nevertheless important quality metric. When the sequence reads are joined together in the wrong order or orientation, some of the exons will be arranged in the wrong order or orientation. This will confuse any gene-annotation program. For example, a 2-kb segment that is flanked by a pair of inverted repeats might be assembled in the wrong orientation. Comparison of independently assembled BACs would not necessarily detect the mistake, because the problem is due to sequence content, not data quality, and the same mistake could be made in both BACs. Comparison with existing physical or genetic maps validate assembly accuracy on the Mb length scale, but that is much larger than the size of most genes. Clone-end pairing information does validate a contig assembly on the kb length scale of the genes. However, when the clone ends are also used to assemble the sequence, they do not qualify as an independent confirmation. To address this problem, we aligned cDNA sequences (i.e., experimentally derived transcripts) with the genome sequence.

We removed obvious redundancies by eliminating any cDNA that was more than 90% contained inside another. Transposon sequences identified by RepeatMasker (*38*), generally in the 3′-untranslated region, were trimmed off to minimize the number of ambiguous hits. Alignments were allowed to span multiple contigs. Within any one contig, a putative misassembly was flagged whenever an exon was missing from the middle of the chain, in the wrong order, or in the wrong orientation. Missing splice sites resulting from minor sequencing errors, and partial alignments resulting from missing sequences at the end of a contig, were not counted. All putative misassemblies were validated by visual inspection, to ensure that no better alignments could be found. If in the end, the best alignment remained problematic, we concluded that there must have been a misassembly. One might think that lower quality cDNA sequences would contribute to the problematic alignments, and that this procedure would only set an upper bound on the number of misassemblies. However, we doubt that this is a serious problem. Substitutional errors might be common in cDNA sequences, but they would not trigger our detection algorithm. Only exon-sized rearrangements, especially those that change the order and orientation, would do so, but such rearrangements are rare in cDNA sequences.

We benchmarked our misassembly detection procedure on two of the most recently completed model organism genomes: *A. thaliana*, which is of finished quality (*4*), and *Drosophila melanogaster*, from the Celera 13× whole-genome-shotgun sequence (*25*). For *A. thaliana*, we detected problems in 0.2% of 4804 genes, and for *D. melanogaster*, we detected problems in 1.1% of 1889 genes. For *93-11* contigs, we detected problems in 1.1% of 907 genes, which was comparable to the *D. melanogaster* data.

Compositional gradients. The rice genome has compositional properties that differentiate it from the other sequenced plant genome, *A. thaliana*, and introduce unique difficulties for genome analysis. Here, we show data on exon, intron, and gene sequences derived from alignment of cDNAs with genomic sequence. Indeed, for Figs. 3 through 7, all of the gene models were derived from cDNA alignments, not gene-pre-

diction programs. GenBank release 125 (15 August 2001) was used for the *A. thaliana* figures, and for the rice cDNAs. The rice genome sequence was our *93-11* assembly. The human cDNA sequence was downloaded on 2 March 2001 from NCBI-RefSeq ftp://ncbi.nlm.nih.gov/refseq/H_sapiens and the human genome sequence was downloaded on 27 February 2001 from ftp://ncbi.nlm.nih.gov/genomes/H_sapiens, immediately after the initial annotation papers.

Genomic, exon, and intron GC contents. The average genomic GC content for prokaryotes and eukaryotes varies widely. It ranges from less than 22% in the human malaria parasite, *Plasmodium falciparum*, to more than 68% in the large amplicon of *Halobactrium* sp. NRC1 (*39*). Local heterogeneity in GC content can be enormous, ranging from 26 to 65% in the human genome alone. In contrast, AG content (purine) is homogeneous (*40–43*), fluctuating by just a few percent about a mean of 50%. Compositional heterogeneity has been debated for more than 30 years (*44–47*). Discussions have focused on the characterization of the human genome as a mosaic of GC-rich and AT-rich "isochores," which are observed in warm-blooded vertebrates, but not in cold-blooded vertebrates. More recently, an elevated GC content in the *Gramineae* (grass) genomes was reported, extending perhaps to all monocot genomes, but not to eudicot genomes (*48*). It is not known whether or how this phenomenon is related to isochores.

Major differences between sequence content in *A. thaliana*, rice, and human are observable even at the simplest level, from distributions of genomic GC content. Traditionally, GC content was computed on a large window size, typically in the 100s of kilobases, to mimic the original Cs_2SO_4 density gradient experiments (*49*, *50*). We have found that smaller windows are more informative, because when these windows are larger than a typical gene size, they obscure differences between intergenic DNA and genes. We used a 500-bp window size, to obtain a smaller size than that of most plant genes (Fig. 2). As previously reported (*51*), the *A. thaliana* distribution displayed a "shoulder" on the AT-rich side, which could be attributed to the sizable fraction of the genome that was in intergenic DNA. The primary peak at 0.382 was nearly identical to the 0.388 GC content of the average *A. thaliana* gene. In contrast, no shoulder was observed in rice. However, a "tail" was apparent on the GC-rich side. The human distribution also displayed no shoulder, but a minor tail might have been present. To analyze these features, we plotted GC content distributions for exons and introns (Fig. 3). Rice exons exhibited a GC-rich tail, but rice introns did not, indicating that the GC-rich tail in the rice genomic distribution was primarily due to the exons.

Variation in GC content within genes. The key question is whether the increase in exon GC content was due to many genes with a few GC-rich exons or to a subset of GC-rich genes. Equivalently, was most of the variation in exon GC content within genes or between genes? After the GC contents of individual exons and introns were plotted as a function of genomic length (i.e., the sum of exon and intron lengths), it was apparent that most of the variation was within genes (Fig. 4). Contrary to the expectation that, in the human genome, large genes are on average more AT-rich than small genes, we found that at least one exon of exceptionally high GC content could be found in almost every rice gene, including the largest ones. Moreover, when the GC content of the protein-coding regions was plotted as a function of position along the direction of transcription, starting

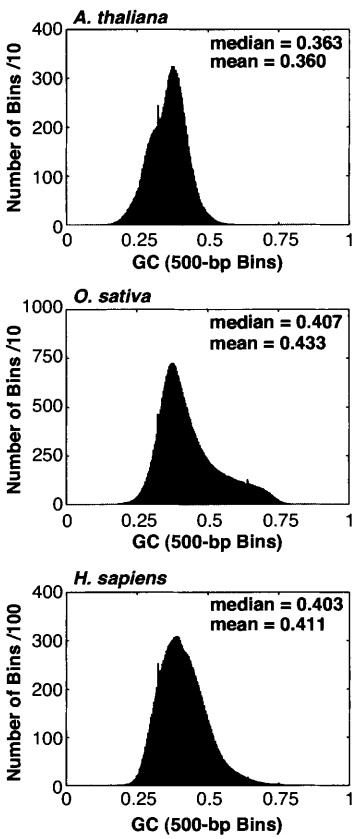

Fig. 2. Distributions for genomic GC content in *A. thaliana*, *O. sativa*, and *H. sapiens*, computed over a bin size of 500 bp. Note that for bins/10 = 100, the number of bins with that GC content is 1000.

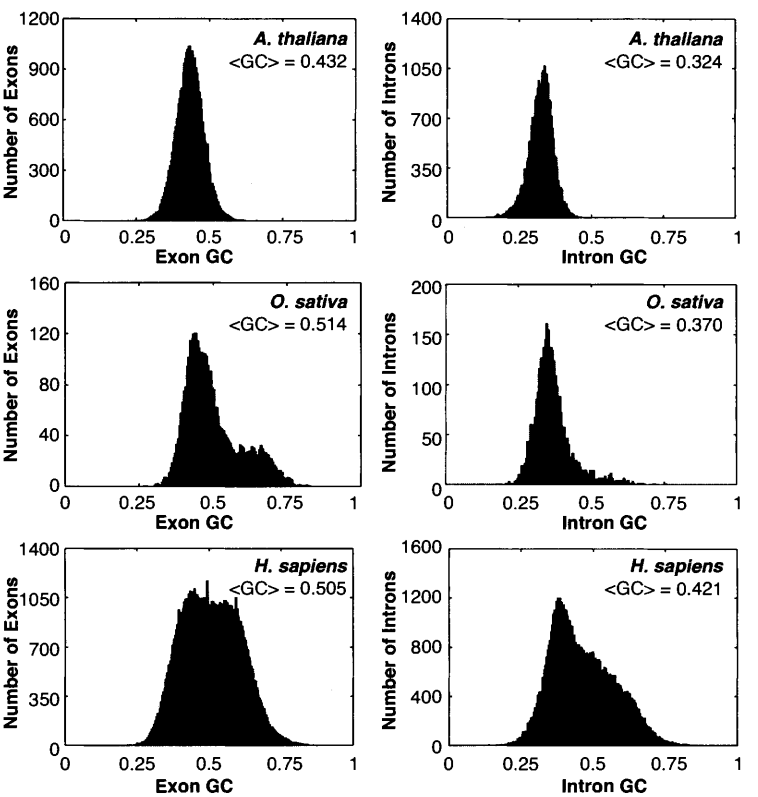

Fig. 3. GC content distribution for exons and introns in *A. thaliana*, *O. sativa*, and *H. sapiens*. All exon and intron sequences were derived from cDNA-to-genomic alignments. Mean GC content is computed on a length-weighted basis as $<GC> = \sum_i L_i \cdot GC_i / \sum_i L_i$, where GC_i and L_i are the GC content and length for the *i*th segment (exon or intron).

from the 5′ end, we observed a negative gradient in the GC content of rice genes (Fig. 5). Typically, the 5′ end was up to 25% richer in GC content than the 3′ end. These gradients would extend to about 1 kb from the 5′ end, before finally petering out. The magnitudes of these gradients varied. A few genes had zero gradients, but almost no genes had positive gradients. In contrast, for *A. thaliana*, no comparable gradients were observed. Examining hundreds of best available homologs (i.e., possible orthologs), we found that the GC content of rice genes was equal to or exceeded that of their *A. thaliana* counterparts, at all positions along the coding region.

A more detailed analysis of this compositional gradient will be presented elsewhere (*52*). The important point is that, not only is there a gradient in GC content, but there are also gradients in the patterns of codon and amino acid usage. The former is a novel challenge for the ab initio gene-prediction programs that rely on codon-usage statistics, and the latter makes it more difficult to do protein homology searches across the monocot-eudicot divide.

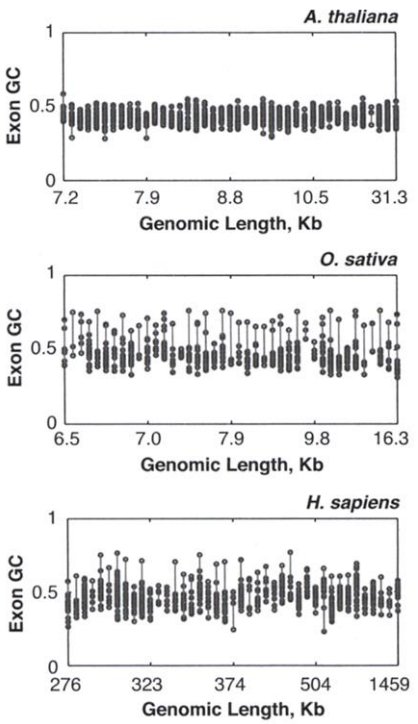

Fig. 4. GC content for individual exons as a function of their gene size, in *A. thaliana*, *O. sativa*, and *H. sapiens*. All exon and intron sequences were derived from cDNA-to-genomic alignments. Each data point is a single exon. Exons for the same gene are plotted at the same abscissa and connected by a vertical line. The genes are sorted by size, where gene size is defined as the sum of exon and intron lengths. To make the figure legible, we use constant spacing between genes, thus resulting in nonuniform abscissa labels. We show only the 41 largest genes for which the entire cDNA could be aligned to genomic sequence. Given the draft nature of the rice genome, some of the largest rice genes had to be omitted.

Genic and intergenic DNA. We examined exon and intron distributions for every plant, vertebrate, and invertebrate organism with more than a hundred or so genes in GenBank, either by cDNA alignments or by parsing annotations. Numerical summaries (Web Supplement 1) are available on *Science* Online at www.sciencemag.org/cgi/content/full/296/5565/79/DC1. Exon sizes are narrowly constrained, but intron sizes can be highly variable within and between organisms. Intron-size distributions tend to be bimodal, weakly (most organisms) or strongly (human). There is always a sharp "spike" at some organism-specific minimum size, which is about 90 bp for plants and vertebrates (Fig. 6). There is also a broad "hump" due to the larger introns. The magnitude of this hump is highly variable between organisms and can be difficult to ascertain precisely, because of the systematic biases against the complete sequencing of large genes.

Although the existence of this acquisition bias is known, the magnitude of its effect on our perception of intron and gene sizes is not well appreciated. For example, in the initial annotation of the human genome, the reported mean gene size of 27 kb turned out to be an underestimate by a factor of 3 (*53*). To correct for this bias, one need only realize that the bias against complete sequencing of large genes is equivalent to the bias against production of large genomic contigs. Thus, the correction can be made by restricting the computation of the mean gene size to cDNA alignments in contigs above a minimum size, and extrapolating to the limit of infinite contigs. For the human genome, the extrapolated mean gene size was 72 kb. In *A. thaliana*, there was no appreciable bias, because most of the contigs were already much larger than the genes, which had a mean size of 2.4 kb. This was larger than the published gene size of 2.0 kb, but only because we included UTRs, whereas the published numbers did not. In rice, there was a small acquisition bias, given the draft nature of our sequence. Nevertheless, the extrapolated mean gene size was only 4.5 kb, much smaller than in the human, consistent with the relatively small hump in the rice intron-size distribution.

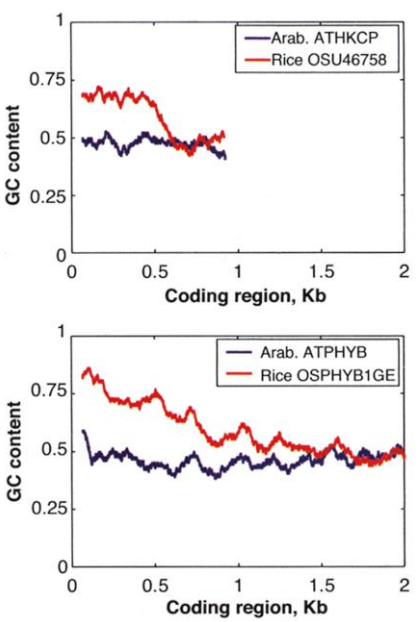

Fig. 5. GC content for homologous genes in *A. thaliana* and *O. sativa* as a function of gene position from the 5′ to 3′ end, computed on a sliding 129-bp window (equal to the median exon size in rice). Only the coding region is shown. GenBank locus identifiers are specified in the legend. The smaller gene is "potassium channel beta subunit," and the larger gene is "phytochrome B."

A preliminary gene count can be estimated from the mean gene size, for comparison against the number of genes identified by the gene-prediction programs. Our estimated 4.5-kb mean gene size for rice is similar to the maximal gene density of one per 4 to 5 kb, based on analyses of syntenic loci across many plant species (*54*). Assuming that the rice intergenic fraction is equal to the 42.2% of the sequence that was in MDRs, and taking 466 Mb as the genome size, the estimated number of rice genes is 59,855. One could also include CpG islands in the gene count, although not every CpG island is associated with a gene, so that this number can at best be considered an upper bound (*55*). Including both assembled contigs and unused reads, 138,485 CpG islands were identified by the standard algorithm (*56*). Either way, rice almost certainly has more genes than *A. thaliana* (*4*), which has only 25,498. It might even have more genes than the human (*7*, *8*), which has 30,000 to 40,000, although the actual gene count remains controversial. The idea that plants might have more genes than humans is not new, as it was predicted before our analysis (*57*, *58*).

Where did the transposons end up? The significance of these size distributions is that a prominent "hump" in the intron-size distribution, as observed for the human, is evidence of extensive transposon activity in the evolution of intron size. RepeatMasker (*38*) identified at least one transposon, and often many more than one, in almost every human intron larger than 1 kb. It rarely found a transposon in smaller introns less than 1 kb, not only in the human, but in every organism analyzed. The negative result might have been due to the incomplete status of the transposon database on which RepeatMasker relies. To support this claim, we introduce an argument that does not rely on knowledge of the sequences of all the extant transposons.

The main assumption was that any transposon should have inserted into the genome

many times before becoming inactive. Despite subsequent degradation of these transposon sequences, portions should remain in many different places throughout the genome. RePS computed the copy number for every 20-mer sequence in the genome, indicating how many times each occurred in the genome. We could therefore determine the copy number required to account for all of a particular sequence data set. These data sets would include all exons, introns, and known transposons (Fig. 7). For the exons and introns, we used cDNA-to-genomic alignments. For transposons, we used RepBase 6.6 (*59*), a database of consensus sequences for every known family or subfamily of transposons. In plants, exons and introns were fully accounted for by 20-mers with copy numbers of less than 10. Transposons required much higher copy numbers of 10 to 10^2 in *A. thaliana* and 10^2 to 10^3 in rice. One could legitimately ask if the absence of large MDR clusters in our rice assemblies was a confounding factor in the intron analysis. We therefore performed an analysis on introns from finished BAC sequences and found no detectable differences. Strikingly, in the human, extremely large copy numbers of 10^4 to 10^5 were required to fully account for the introns, as was observed with the transposons. Human exons, however, were found at the same low copy numbers as in plants.

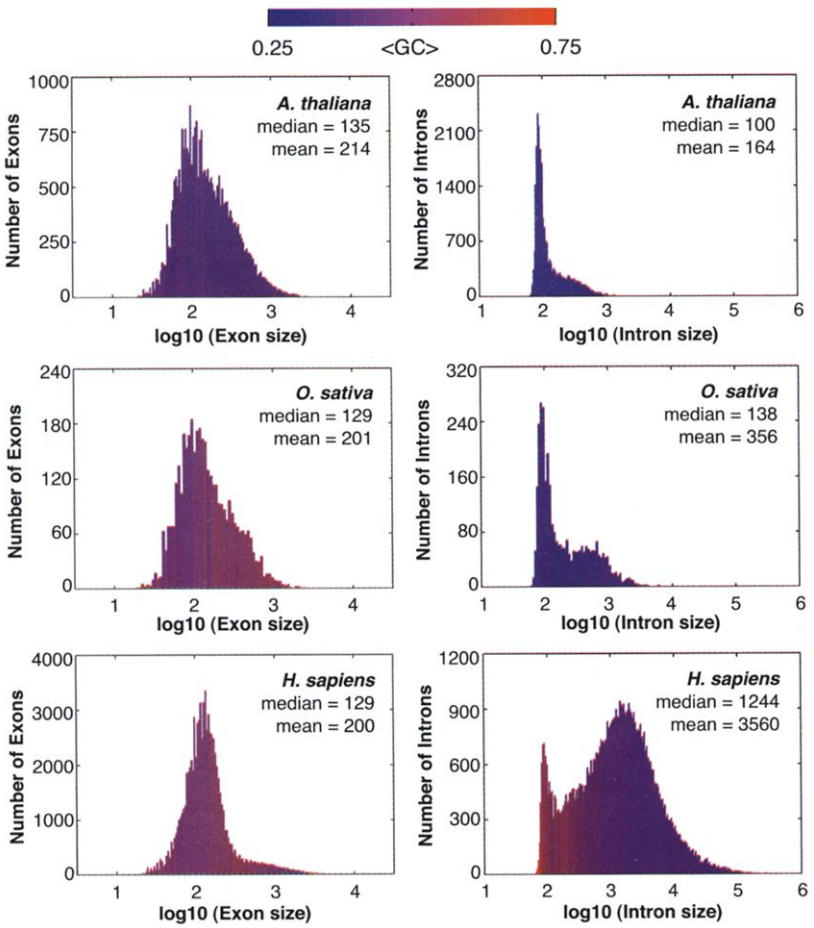

Fig. 6. Exon- and intron-size distributions for *A. thaliana*, *O. sativa*, and *H. sapiens*, with color indicating averaged GC content for exons or introns at that size range. All exon and intron sequences were derived from cDNA-to-genomic alignments.

The copy number analysis shows that few plant transposons are in the introns, and by definition, plant transposons must be located in the intergenic regions between genes. Conversely, analyses of gene size show that most human transposons are in the introns (*53*). We believe that this dichotomy in where the transposons ended up reflects a fundamental difference in plant and vertebrate genomes. The dichotomy is not due to any lack of transposons in plants, because plant genomes contain many transposons. At least 24.9% of the rice genome was identifiably of transposon origins, based on a weighted average of assembled contigs and unused reads, but the correct percentage is likely to be much higher, because the transposon databases on which RepeatMasker relied were incomplete. *A. thaliana*, being a more compact genome, had a reported transposon fraction of 10%, although we suspect that this too is an underestimate.

Repetitive sequences. We deal with three classes of repeats: simple repeats [e.g., $(CAG)_n$], complex repeats (i.e., transposable elements or TEs), and mathematically defined repeats (MDRs). Here, we focus on the first two classes, which we called biologically defined repeats (BDRs). As with intron and gene sizes, acquisition biases must be factored in, so that we do not introduce additional discrepancies among the published studies. For example, a survey of 73,000 sequence-tagged connectors, totaling 48 Mb of sequence from *japonica* (*60*), found that 63% of identified TEs were retrotransposons (e.g., *copia* and *gypsy*). However, a survey of 910 kb of rice genomic sequence (*61*) found that 18.6% of identified TEs were retrotransposons. Most of the remainder were MITEs,

Table 2. Simple repeats. Shown are tandem repeats with periods 1 to 4 (mono-, di-, tri-, and tetranucleotide) and the totality of repeats with all periods. The index *n* is the number of periodic units. For example, AGTTAGTT is a tetranucleotide of $n = 2$. We compute mean GC contents of the observed repeats in each category. Repeat content is then given as a percentage by length, normalized with respect to the data set (assembled contigs, fully masked reads, or cDNAs).

	93-11 assembled contigs				*93-11* fully masked reads				Full-length cDNAs			
	$n = 6$–11		$n > 11$		$n = 6$–11		$n > 11$		$n = 6$–11		$n > 11$	
	% GC	% of data set	% GC	% of data set	% GC	% of data set	% GC	% of data set	% GC	% of data set	% GC	% of data set
Mononucleotides	7.63	1.7847	27.65	0.0680	20.34	0.6953	21.02	0.0154	24.08	0.6303	1.31	0.7125
Dinucleotides	35.77	0.0904	13.08	0.0847	41.86	0.0553	4.38	0.0294	46.85	0.0573	31.11	0.0394
Trinucleotides	71.79	0.0454	10.08	0.0106	67.20	0.0098	20.81	0.0012	83.05	0.1335	66.67	0.0043
Tetranucleotides	28.77	0.0072	24.90	0.0032	37.35	0.0020	31.90	0.0010	50.00	0.0018	0.00	0.0000
All periods		**1.9277**		**0.1665**		**0.7624**		**0.0469**		**0.8229**		**0.7561**

or miniature inverted-repeat TEs, which accounted for 71.6% of identified TEs. Similarly large discrepancies were encountered in analyses of microsatellite distributions using mixed data from BAC ends, ESTs, and finished BAC/PACs (*62*).

For the *93-11* sequence, it is particularly important that we analyze those sequence reads that were not assembled into contigs. Tables 2 and 3 thus summarize repeat contents in the two largest components: 361 Mb of assembled contigs and 78 Mb of unused fully masked reads. Weighted averages for the entire rice genome were also computed. For comparison, we show repeat content in 907 nonredundant full-length cDNAs from GenBank release 125 (15 August 2001). Absolute numbers are not listed because, with so much of the genome in unassembled reads, and with so many of the transposons nested inside some other transposon, accurate counts were not feasible. Results are listed as a fraction of total sequence length.

Simple sequence repeats (SSRs). SSRs are particularly useful for developing genetic markers. They are believed to vary through DNA replication slippage (*63–65*), and are related to genetic instability (*66*). In Table 2, we describe SSR content for two sectors, n = 6 to 11 units and $n > 11$ units, to emphasize that the number of SSRs dropped substantially after 11 units. The SSR content for *93-11* was 1.7% of the genome, lower than in the human, where it was 3% (*7*). The overwhelming majority of rice SSRs were mononucleotides, primarily $(A)_n$ or $(T)_n$, and with n = 6 to 11. In contrast, for the human, the greatest contributions came from dinucleotides. Notably, trinucleotides with n = 6 to 11 were a barometer of gene content. The basic effect was captured by the ratio of trinucleotide to dinucleotide content, which was 2.33, 0.50, and 0.18 in cDNAs, assembled contigs, and fully masked reads, respectively. As required for a barometer, these numbers are well correlated with presumed gene content. In addition, the GC content of these trinucleotides was high, consistent with the high GC content of many rice exons.

Complex sequence repeats (TEs). Transposons identified by RepeatMasker (*38*) were assigned into three classes. Class I repeats are retrotransposons, primarily *Ty1/Copia*-like and *Ty3/Gypsy*-like. Class II repeats are DNA transposons, including *Ac/Ds*, *En/Spm*, *Mariner*-like, and *Mutator* elements. Class III repeats are a previously unknown type of short DNA transposons called MITEs (*67*, *68*). The two common examples are *Stowaway* and *Tourist*. Recently, an active family of *tourist-like* MITEs was identified in maize (*69*). Programs like RepeatMasker identify sequences that share at least 50% identity with a known TE. Because TEs are under no selective constraints after they insert in a genome, they tend to diverge from their ancestral sequence, and become unrecognizable over a time scale of a hundred million years (*70*). Identifiable repeat content is thus a function of TE age and completeness of the TE databases. The numbers listed in Table 3 must therefore be considered underestimates.

Fully masked reads were composed of 59% identifiable TEs. Assembled contigs were only 16%. Of these TEs, the amount in class I and class III repeats was 97 and 1%, respectively, for fully masked reads, but 42 and 40% for assembled contigs. This extremely biased distribution is notable, because class I repeats reportedly inhabit the intergenic regions (*23*, *24*), and class III repeats are found near, although not necessarily in, the genes (*71*). Thus, we had 92.0% functional coverage despite having only 361 Mb in assembled contigs, in a genome of total size 466 Mb. The reason class I repeats failed to assemble is apparent when one examines their mean size. Class III repeats were usually smaller than 671 bp, but class I repeats were as large as 7 kb. Our ability to close repeat-masked-gaps, or RMGs, was limited by the clone-insert sizes. For this assembly, the clone-insert sizes were only 2 kb, although we plan to use larger sizes for the next stage of the rice genome project.

Finally, the TEs in rice cDNAs constituted only 1% of the sequence, which is much lower than the 4% that was reported for human genes (*72*). Gene-associated TEs, in human and other vertebrates, have been proposed to play crucial roles in creating new genes (*73*) and in changing the regulatory circuitry to promote evolution in the host genome (*74*).

Rice gene annotations. Gradients in GC content and codon usage for rice genes create special problems in the gene-annotation process (*52*). Because rice genes have different compositional properties at their 5′ and 3′ ends, it is difficult to train a program to perform well under all circumstances. Some ab initio gene-prediction programs can use different codon-usage statistics for different genes, on the basis of regional GC content, but none use different codon-usage statistics at different positions along the same gene. Unless the gradient is explicitly modeled, or perhaps, codon-usage statistics are abandoned altogether, performance will be subject to the vagaries of the training process. With this in mind, we set out to survey all of the programs trained for rice: FGeneSH (*75*), GeneMark (*76*), GenScan (*77*), GlimmerM

Table 3. Complex repeats. Transposons identified by RepeatMasker are assigned to three classes. Each class has a number of families (e.g., *tourist*-like MITEs), and each family has a number of different subfamilies. The number of subfamilies is listed, as well as their total and mean size. Repeat content for each family is given as a percentage by length, normalized with respect to the data set (assembled contigs, fully masked reads, or cDNAs) or with respect to all identified transposons.

		Number	Total (bp)	Mean (bp)	*93-11* assembled contigs		*93-11* fully masked reads		Full-length cDNAs	
					% of data set	% of repeats	% of data set	% of repeats	% of data set	% of repeats
Class I	LINEs	5	18,997	3,799	1.1905	7.43	0.1318	0.22	0.0257	2.51
	SINEs	7	1,254	179	0.0888	0.55	0.0047	0.01	0.0268	2.61
	gypsy-like	19	105,614	5,559	3.7285	23.28	41.6894	70.35	0.1238	12.07
	copia-like	5	35,151	7,030	1.7175	10.72	15.8506	26.75	0.0869	8.47
	Subtotal				**6.7254**	**41.99**	**57.6766**	**97.33**	**0.2631**	**25.65**
Class II	*Ac/Ds* TEs	3	1,567	522	0.1099	0.69	0.0145	0.02	0.0000	0.00
	En/Spm TEs	3	5,558	1,853	0.2590	1.62	0.2770	0.47	0.0000	0.00
	MULEs	22	25,800	1,173	2.4500	15.30	0.6378	1.08	0.1807	17.62
	Subtotal				**2.8190**	**17.60**	**0.9293**	**1.57**	**0.1807**	**17.62**
Class III	Stowaway-like	70	16,112	230	2.2370	13.97	0.1247	0.21	0.1910	18.62
	tourist-like	77	19,933	259	3.7405	23.35	0.3228	0.54	0.3451	33.65
	Unknown MITEs	2	1,341	671	0.4950	3.09	0.2080	0.35	0.0458	4.46
	Subtotal				**6.4725**	**40.41**	**0.6556**	**1.11**	**0.5818**	**56.73**
	Grand Total	**213**	**231,327**	**1,086**	**16.0169**	**100.00**	**59.2615**	**100.00**	**1.0255**	**100.00**

(*78*), and RiceHMM (*79*). Strictly speaking, GenScan was trained for maize, another monocot with GC content gradients.

Assessment of gene-prediction programs. All the gene-prediction programs were pretrained by the authors and tested against our cDNA-to-genomic alignments. These comparisons may favor the program that was trained on the largest and most recent data set, but that information was not available to us. Performance was measured at the base pair and the exon levels, and then plotted as a function of position from 5′ to 3′ end (Fig. 8). Sensitivity is the probability that the actual coding region is correctly predicted (1 minus false-negative rate). Specificity is the probability that the predicted coding region corresponds to the actual coding region (1 minus false-positive rate). Some programs, including GenScan, had sensitivities that were extremely dependent on position, although this was not the case when we applied these performance metrics to human genes. This suggests that the compositional gradients were indeed a source of error. That GenScan would be affected is significant because, in the most recent comparative analysis of human genes (*80*), two of the most successful programs were FGeneS (a variant of FGeneSH) and GenScan. For rice, however, FGeneSH is the most successful program. It is not obvious why, although the documentation states that FGeneSH places more weight on signal terms (e.g., splice sites, start and stop codons) than on content terms (i.e., codon usage).

Submitting our *93-11* assembly to the FGeneSH Web site returned 75,659 predictions. However, only 53,398 were complete, in the sense that initial and terminal exons were both present; 7489 had only an initial exon, 11,367 had only a terminal exon, and 3405 had neither. When we include predictions without both an initial and terminal exon as only half a gene, we obtain an upper bound of 64,529 genes. Without correcting for sensitivity or specificity, the estimated gene count is 53,398 to 64,529. This is similar to the 59,855 genes that we predicted from considerations of gene size and repeat content. How good are these predictions? We have reservations about the absolute value of the performance metrics, because FGeneSH was probably trained on a gene set with considerable overlap to our reference cDNAs. These metrics may not tell us how well FGeneSH performs for rice genes with substantially different compositional properties. However, their relative values should be interpretable. Namely, base-level specificities were better than base-level sensitivities, indicating that false-negatives are more likely to be a problem than false-positives. The program is more likely to miss an exon fragment than to label something part of an exon by mistake. Sensitivities and specificities were much worse at the exon level, implying that that the exon-intron boundaries are not precisely defined, even when the presence of a gene is correctly detected.

Two pieces of evidence qualify our level of confidence in the gene predictions. First, if the sensitivity is really as good as suggested, then we ought to be able to find most of the ESTs in the predicted gene set. We thus performed a comparison against the 24,776 UniGene clusters assembled from our 87,842 ESTs. The result was that only 77.3% of these clusters could be found in the FGeneSH predictions. Second, the mean size of the predicted coding regions in rice was only 328 residues, or 73.5% of the predicted coding regions in *A. thaliana*, which averaged 446 residues. This was the case even though we restricted the mean to complete genes with initial and terminal exons. Although it is possible that rice genes are intrinsically smaller than *A. thaliana* genes, we believe that this discrepancy reflects a deeper problem that is related to the compositional gradients, as will be explained below.

Functional classification of rice genes.

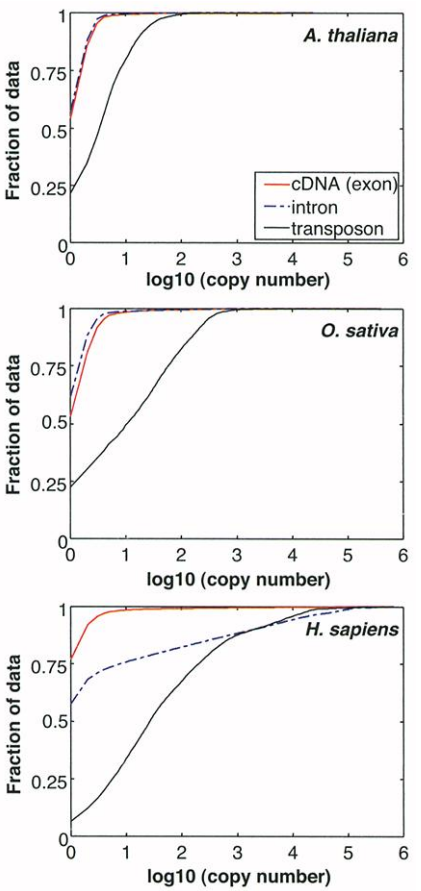

Fig. 7. Cumulative copy numbers for exons, introns, and known transposons in *A. thaliana*, *O. sativa*, and *H. sapiens*. We determined the copy number of 20-mers in each genome, and then mapped these 20-mers back to exons, introns, and known transposons for each genome. All exon and intron sequences were derived from cDNA-to-genomic alignments. The analyzed transposons were the consensus sequences for the known families or subfamilies of transposons. We show here the fraction of each data set that is in 20-mers up to the indicated copy numbers.

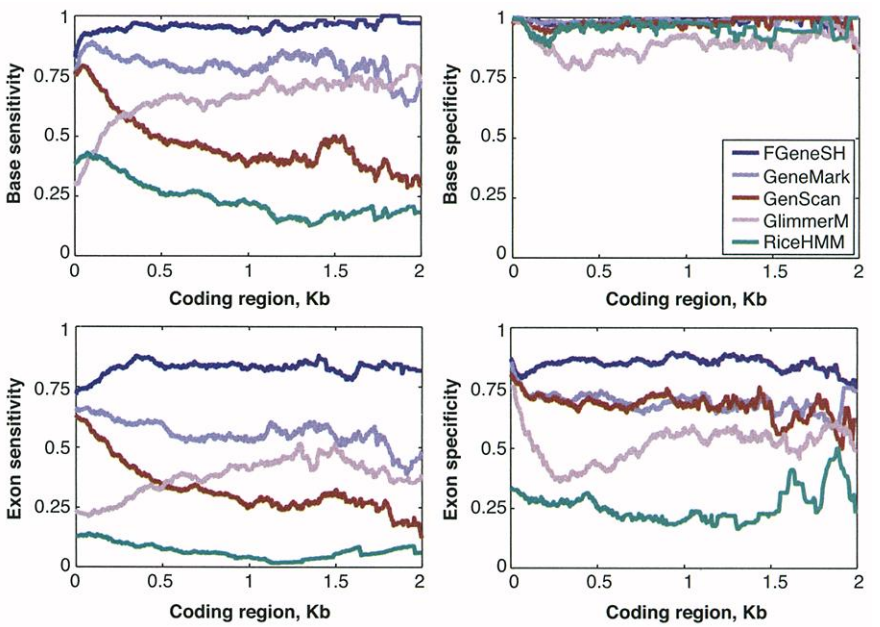

Fig. 8. Performance metrics for ab initio gene-prediction programs, as a function of gene position from 5′ to 3′ end, when compared against cDNA-to-genomic alignments at the same loci. Sensitivity is the probability that the coding region is correctly predicted (1 minus false-negative rate). Specificity is the probability that the predicted coding region is real (1 minus false-positive rate). At the exon level, both splice sites must be correctly predicted for an exon to be counted as correct.

Although 25,426 genes have been identified in *A. thaliana*, fewer than 10% have been documented experimentally (*81*). Consequently, functional classification of plant genes must rely heavily on homology, coupled with a few nonhomology-based methods, such as phylogenetic profiling, correlated gene expression, and conserved gene orders. Only 27.3 and 36.3% of *A. thaliana* genes have been classified by InterPro (*82*) and Gene Ontology Consortium (*83*), respectively. To establish functional classifications for rice genes, we performed protein-to-protein sequence comparisons against *A. thaliana* annotations, and adopted classifications from the best match to *A. thaliana*. We considered only those 53,398 predictions from FGeneSH with initial and terminal exons. When multiple hits were found, we selected the one with the longest extent of homology. We required that at least 25% of the protein length be matched. This is a low-threshold setting, but as we will explain below, it was necessary. In total, 15.9 and 20.4% of rice gene predictions were classified by InterPro and Gene Ontology Consortium, respectively. As a percentage of classified genes, the predicted gene sets for rice and *A. thaliana* are similarly distributed among different functional categories (Fig. 9). We depict Gene Ontology Consortium because more genes were classified. Tables of predicted rice genes and their functional classifications (Web supplement 2), as well as InterPro figures (Web supplement 3), are available on *Science* Online at www.sciencemag.org/cgi/content/full/296/5565/79/DC1.

We advise extreme caution in interpreting minor differences in functional classification between the predicted gene sets for rice and *A. thaliana*. With such a large fraction of the genes unclassified, intrinsic uncertainties in any classification scheme are amplified into artifactual differences. For example, the largest difference for InterPro was in signal transduction genes, but no notable difference was observed for Gene Ontology Consortium. Furthermore, focusing on small differences that had a high likelihood of being artifactual would distract from the major difference between rice and *A. thaliana*, which as we will show next, lies almost entirely in those genes with no functional classification.

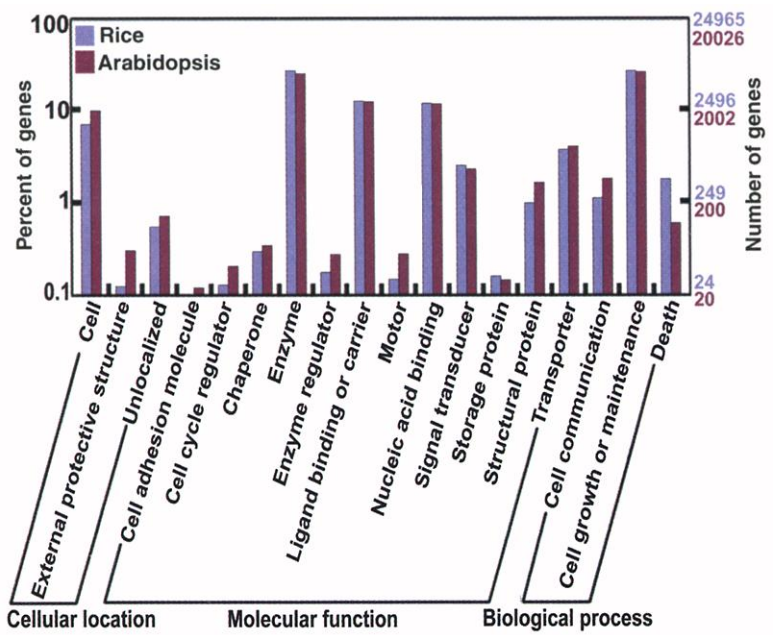

Fig. 9. Functional classification of rice genes, according to Gene Ontology Consortium, and assigned by homology to categorized *A. thaliana* genes. In this ontology, "biological process," "cellular location," and "molecular function" are treated as independent attributes. Only 36.3% of the 25,426 predicted genes for *A. thaliana* are classified. For rice, only 20.4% of the 53,398 complete predictions, with both initial and terminal exons, could be classified.

Table 4. Polymorphism rates relative to *93-11* (*indica*). Comparisons were made to finished BAC sequences from *GLA* (*indica*) and *Nipponbare* (*japonica*), as well as to *PA64s* contigs. Rates were computed for repeated and unique regions, in single-base substitutions (SNPs) and insertion-deletions (InDels). The numbers given for "unalignable" are a gross underestimate because RePS assemblies omit many of the fully masked reads that correspond to the unalignable regions of Fig. 14.

	Nipponbare (*japonica*)	*PA64s*	*GLA* (*indica*)
SNPs in repeated sequence (%)	0.88	0.68	0.65
InDels in repeated sequence (%)	0.33	0.45	0.27
SNPs in unique sequence (%)	0.50	0.35	0.50
InDels in unique sequence (%)	0.14	0.16	0.15
Repeated sequence fraction (%)	24.1	25.5	22.8
Unique sequence fraction (%)	74.8	74.3	74.1
Parts unalignable by BLAST (%)	1.1	0.3	3.1

***A. thaliana* comparisons.** In general, there are two ways to compare gene sets: through colinearity and homology. Colinearity of plant genomes has been studied extensively (*84*, *85*). For analyses done within a plant family, high degrees of colinearity have been consistently observed. Across the monocot-eudicot divide, with rice and *A. thaliana* as representative species, observed degrees of colinearity have been considerably lower (*86–89*). For example, an analysis of a 340-kb segment on rice chromosome 2 identified 56 putative genes (*88*). Homologs for 22 (39%) of them were identified in *A. thaliana*, but were distributed among 5 chromosomal segments, with several small-scale inversions. Another study of 126 rice BACs, totaling 20 Mb of sequence and with 3011 putative genes, identified homologs in *A. thaliana* for 1747 (58%) of these genes (*89*). Typically, each 150-kb BAC mapped to three or more chromosomes. Notwithstanding the absence of colinearity, the finding that only half of the rice genes had a homolog in *A. thaliana* was unexpected. Although these analyses were based on predicted genes, which have not yet been confirmed, we do not believe that this was why so few rice genes had a homolog in *A. thaliana*, because a similar analysis was done with 27,294 unique ESTs from *Z. mays* (maize), and only 62% of the open reading frames had a homolog in *A. thaliana* (*90*).

We focus exclusively on homology, rather than orthology, because extensive gene duplications in *A. thaliana* (*4*, *91*) and rice make strict one-to-one pairing relations, the classic definition for orthology (*92*), difficult to determine. A mere 35% of *A. thaliana* genes are unique and 37.4% belong to gene families with more than five members. Segmental duplications larger than 100 kb in size constitute 58% of the genome, and 17% of the genes are arranged in tandem arrays. In comparisons of rice with *A. thaliana*, and vice versa, we sought to compute the degree of homology in each direction, and the extent to which gene duplications in *A. thaliana* are replicated in rice when decomposed by functional classification. Even this modest objective was not easy to accomplish, because of unexpected complications intro-

duced by the compositional gradients in rice.

Homology between monocots-eudicots. The complete set of 25,426 annotated *A. thaliana* genes was downloaded from the *Arabidopsis* Information Resource Web site (*93*) on 29 November 2001. As a control, 1441 proteins were downloaded from SwissProt (*94*) on the same day. The rice genes were restricted to the 53,398 predictions from FGeneSH with initial and terminal exons. We compared protein sequence to all six reading frames of the genome sequence by means of TblastN (*36*). Therefore, if the homology search failed, it would not be due to a gene being missing from the annotation of the target genome. The expectation value cutoff was set to 10^{-7}. This was not a sensitive parameter, as most hits were either very good or very bad. What mattered was the "coverage rule." We projected every hit back to the protein query, and unless a minimum fraction of the protein was covered, none of the hits were accepted. The hits had to occur in the same order in both the query and the target, and they all had to be in the same orientation. When a homolog spanned more than one scaffold, the coverage rule was imposed on each scaffold. From this rule, we estimated the number of homologs per gene, the extent of the homology, and the percentage amino acid identity (*95*).

The asymmetry in the monocot-eudicot analysis was striking (Fig. 10). About 80.6% of *A. thaliana* genes had a homolog in rice. The mean extent of homology was 80.1% of the protein length, and there was 60.0% amino acid identity. If instead of the full set of annotated genes, we had used SwissProt genes, 94.9% of the genes would have had a homolog, across 86.7% of the protein length and at 72.9% amino acid identity. Presumably, there were more homologs in the SwissProt data because they were more biased toward highly conserved proteins. In contrast, only 49.4% of predicted rice genes had a homolog in *A. thaliana*. The mean extent of homology was 77.8% of the protein length, and there was 57.8% amino acid identity. For brevity, predicted rice genes with a homolog in *A. thaliana* are called WH genes, and those with no homologs are called NH genes. We identified two distinct problems in this analysis, both attributable to the compositional gradients in rice. One was the poor quality of the FGeneSH predictions for NH genes, and the other was related to the probability of identifying a TblastN hit even with a perfect gene annotation. We did use ESTs to confirm that NH genes were not false predictions, but first, we will discuss what we believe to be the true problems.

We had previously observed that rice gene predictions were only 73.5% the size of *A. thaliana* gene predictions. This discrepancy is not due to the WH half of the rice genes. It is due to the NH half, which was on average 49.4% smaller than the WH half (Fig. 11). To analyze the problem, we randomly sampled 3000 WH genes and 3000 NH genes, and applied the analyses of Fig. 3 to Fig. 7. In general, WH genes resembled the "gold standard" based on alignment of cDNA to genomic sequence. NH genes exhibited a number of striking differences. First, the decreased coding region size was clearly due to a decrease in the number of exons, not to a decrease in the size of the exons. The GC-rich tail in NH gene exon distribution was twice as large as normal (Fig. 11), suggesting that NH genes had more pronounced GC content gradients than either WH genes or those cDNAs retrieved from GenBank. It is plausible that FGeneSH performance would have faltered on NH genes, because NH genes did not resemble those genes on which FGeneSH was presumably trained. NH genes also had twice as many introns as normal in the 200- to 2000-bp range (Fig. 11). This would be consistent with some of these missing exons being combined with their flanking introns. The preponderance of anomalous subminimal introns would be consistent with exon fragments being mistakenly called introns. However, NH genes could not be transposon sequences, because a 20-mer analysis confirmed that their constituent sequences were found in the genome at low copy numbers, much like WH and cDNA-derived genes.

Although we do not entirely ascribe the small size of NH genes to a failure by FGeneSH to detect exons, it is likely that more exons were missed than for WH genes. Thus, it would be more difficult to identify a homolog for these genes in *A. thaliana*. However, even for experimentally derived gene

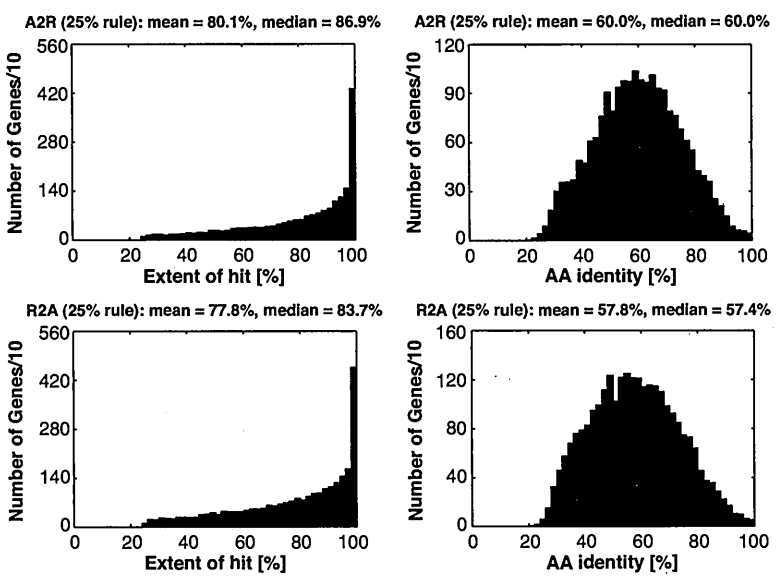

Fig. 10. Distributions in extent of homology and maximum amino acid identity, for *Arabidopsis*-to-rice and rice-to-*Arabidopsis* comparisons. These values are based on a comparison of predicted protein sequence against all six reading frames of the target genome sequence.

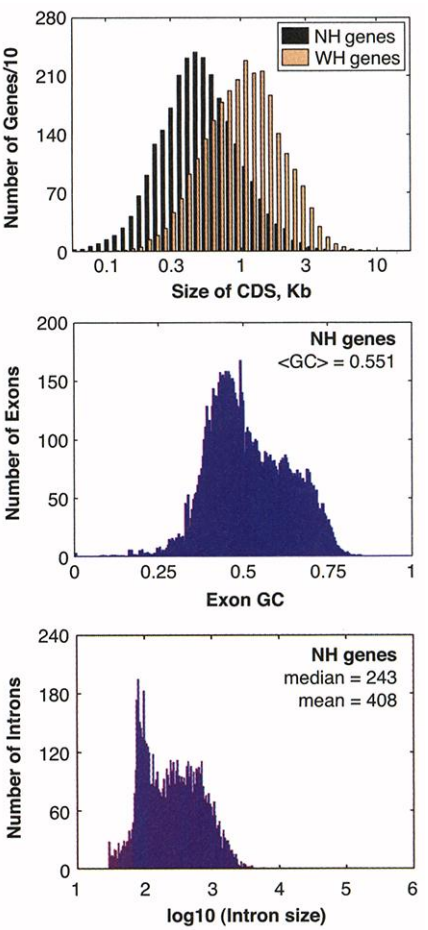

Fig. 11. Size distribution of predicted rice genes with a homolog (WH), and with no homolog (NH), in *A. thaliana*, plus exon GC content and intron size for a random sampling of 3000 NH genes. Gene size refers to the size of the predicted coding region.

sequences, like cDNAs, the probability of identifying a TblastN hit, as a function of the position, dropped precipitously near the 5′ end of the genes (*52*). Far from the end, the probability was about 90%, but within the first few hundred bases near the 5′ end, the probability dropped to less than 50%. This was another consequence of the compositional gradients in rice. The magnitude of the effect was unexpected. We had thought that selective constraints on coding sequences would have limited the number of amino acid changes, despite pressure from rising GC content. However, this was not the case. Homology searches were more likely to fail with the smaller NH genes because the problematic region was a larger fraction of their total length, and our "coverage rule" required that the TblastN hits cover a minimum fraction of the coding region. Even in the *Arabidopsis*-to-rice analysis, where the gene predictions were more reliable, 83.2%, 80.6%, 69.5%, and 48.5% of *A. thaliana* genes had a homolog in rice, for coverage rules of 0, 25, 50, and 75%. We had to use a relatively low coverage rule of 25%. Given the typical protein and protein domain sizes of 446 and 100 residues (*96*–*98*), respectively, this was equivalent to one protein domain.

Alternatively, what if the problem were due to scaffold size? Half of the NH genes were identified in a scaffold that was smaller than 7.1 kb. However, as a function of scaffold size, predicted coding regions for NH genes were almost always the same size. NH genes found in scaffolds greater than 7.1 kb were only 7% larger than those found in scaffolds less than 7.1 kb. Scaffold size could not have been responsible for the small size of the NH genes. Perhaps NH genes are not real genes at all. Are they even expressed? Looking back at our EST confirmation analysis, we found that 42.9% of WH genes were confirmed by a UniGene cluster, compared with 15.4% of NH genes. Assuming that all WH genes are real, this would imply that (15.4/42.9) × 100% = 35.9% of NH genes are real. However, if we adjust for their being 49.4% smaller than normal, attributing this size deficit to missed exons, then 72.7% of NH genes are real. Certainly, not every NH gene is real, but many are. To be conservative, we can adjust our gene count estimates by a factor of (0.494 + 0.727 × 0.506), resulting in a revised gene count of 46,022 to 55,615.

Considering the relatively recent divergence between monocots and eudicots, 145 to 206 million years ago, it is surprising to find so many genes in rice with no homolog in *A. thaliana*. Even more intriguing, this absence of homology for NH genes extended to other sequenced organisms, including *D. melanogaster*, *Caenorhabditis elegans*, *Saccharomyces cerevisiae*, and *Schizosaccharomyces pombe*. Although WH genes had a 30.5% probability of being homologous to at least one gene in one of these organisms, NH genes had a 2.4% probability. Hence, the major difference between rice and *A. thaliana* gene sets lies in that half of the predicted rice gene set with essentially no homologs in any organism, and whose functions are largely unclassifiable.

Duplication between monocots-eudicots. Having established the major difference between the gene sets for rice and *A. thaliana*, we now consider the similarity. We had reported that 80.6% of the predicted *A. thaliana* genes, and 94.9% of the SwissProt genes, had a homolog in rice. The actual number is likely to be even higher, because the gradients kept us from identifying potential homologs for smaller genes. We know that, within *A. thaliana*, the genes are highly duplicated. Are these genes duplicated in the same manner when mapped to rice? As a proxy for the number of gene homologs within and between genomes, we used the "hits per gene," as defined in the notes (*95*). Considering that, in the *Arabidopsis*-to-rice comparison, we used a low coverage rule of 25% to compensate for the gradients, it was inevitable that we would experience more difficulty than usual in distinguishing between duplicated domains and duplicated genes. Thus, the number of hits per gene is an overestimate of the number of gene homologs.

Comparing *Arabidopsis*-to-*Arabidopsis* (A2A), the mean and median hits per gene were 38.2 and 6.0, similar to the mean and median of 33.4 and 5.0 that we observed comparing *Arabidopsis*-to-rice (A2R) (Fig. 12). That the A2R numbers would be slightly smaller makes sense, given the 145 to 206 million years of divergence. We further note that the means were large only because of a few outliers, some with up to 1000 hits. The identity of these outliers included protein kinase, cytochrome P450, putative disease resistance, and many "unknown" genes. It is difficult to draw any conclusions about the last category, but the others are highly duplicated gene families, which confirms that these outliers were not computational artifacts. The maximum amino acid identity was independent of the number of hits, but the minimum amino acid identity decreased with the number of hits, which would be consistent with an increasing occurrence of hits to ever larger families of related but divergent genes. Although the number of hits was dependent on the functional classification, it was similarly distributed among the different functional categories for A2R and A2A (Fig. 13). Therefore, not only was it possible to identify a homolog in rice for almost every *A. thaliana* gene, but the patterns of gene duplication in one were largely replicated in the other.

The most parsimonious explanation is that the rice gene set is essentially a "superset" of the *A. thaliana* gene set. However, we are unable to say how many of these additional genes that are unique to rice are functionally novel, or merely unrecognizable, because of gradients in rice amino acid usage. It does seem unlikely that so many novel genes would arise within only 145 to 206 million years, and therefore, we suspect that a massive duplication event (or a series of duplication events) occurred, after which many of the rice genes were rendered unrecognizable

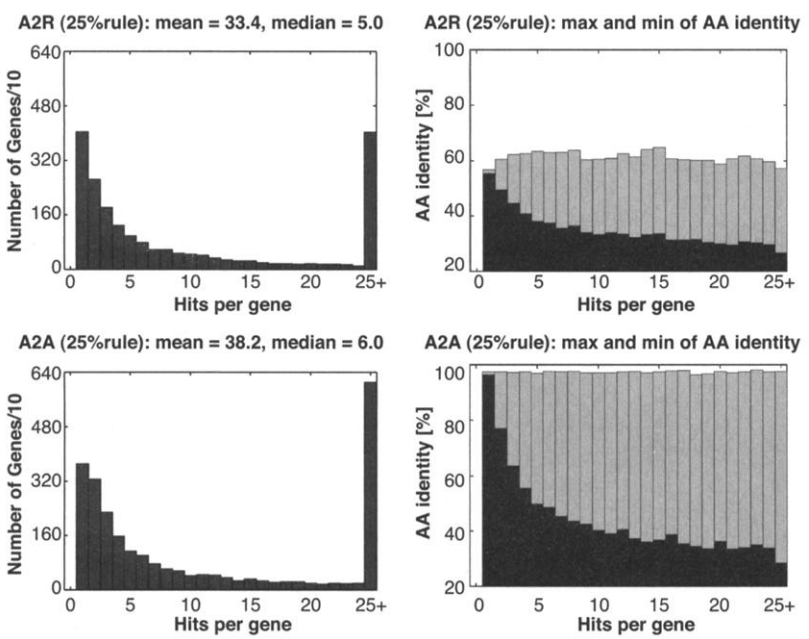

Fig. 12. Distributions in number of hits per gene and maximum-versus-minimum amino acid identity, for *Arabidopsis*-to-rice and *Arabidopsis*-to-*Arabidopsis* comparisons. "Hits per gene" is a proxy for the number of gene homologs, between and within genomes.

by compositional gradients. Some may have been inactivated, and now exist only as pseudogenes. However, until we can compensate for the confounding effects of compositional gradients, we cannot explore the extent to which rice (*99*) and many other plants, including *A. thaliana*, are hybrid (*100*) or allopolyploid (*101*, *102*) in origin.

Rice polymorphisms. Differences between subspecies or cultivars of rice must be described at two levels, gross and nucleotide. At the gross level, we found kilobase-sized regions of high similarity interspersed with kilobase-sized regions of no similarity. One such example is shown in Fig. 14, which was based on a comparison of two overlapping BACs from *indica* and *japonica*. Every unalignable region coincided with a cluster of MDRs, traceable to length differences of 0.7 to 25 kb between the two source sequences, distributed in almost equal proportions between insertions and deletions. To the extent that BDRs could be identified, in roughly half of the unalignable regions, they belong to the class of nested retrotransposons that inhabit the intergenic regions between genes. This is another confirmation of the observation that genome sizes change rapidly in grasses (*103*). On the basis of the available 259 kb of overlapping finished BAC sequences, from *Nipponbare* (*japonica*) and *GLA* (*indica*), all on rice chromosome 4, we would estimate that 16% of the *indica* and *japonica* genome is unalignable by this definition.

At the nucleotide level, excluding the unalignable regions, we define polymorphism rates for repeated and unique sequence, partitioned in single-base substitutions (single-nucleotide polymorphisms, SNPs) and insertion-deletion polymorphisms (InDels). By repeated sequence, we mean MDRs. Three different comparisons are shown in Table 4. Two are based on the alignment of *93-11* contigs to finished BAC sequences from *Nipponbare* (*japonica*) and *GLA* (*indica*), totaling 11.8 and 0.9 Mb, respectively. The other is a comparison of *93-11* and *PA64s* contigs. One might question the accuracy of a polymorphism rate based on rough draft sequence, particularly the low-coverage *PA64s* sequence. However, as we noted in our "quality assessments" section, most of the errors are in the small contigs and at the ends of the contigs. Thus, we restricted this analysis to contigs larger than 3 kb, with 500 bp trimmed off both ends. Overall, there was twice as much variation in the repeated regions as in the unique regions. Substitution rates were two to three times as large as InDel rates. Remarkably, there was very little difference among the three pairwise comparisons. For *93-11* to *PA64s*, averaged over repeated and unique regions, the SNP and InDel rates were 0.43 and 0.23%, respectively. Combining the SNP and InDel rates, we obtained an overall rate of 0.67%. Although the numbers are not exactly comparable, the measured polymorphism rate in maize was 0.96% (*104*).

SNPs are useful in genetic mapping (*105*), and are either directly applicable to phenotypes or indirectly applicable through linkage and association studies. Polymorphisms in the unique regions are particularly useful because, unlike those in the repeated regions, they are more reliably genotyped with existing high-throughput technologies, which always involve some sort of hybridization step. We expect that genome-wide SNP mapping in plants (*106*) will become more popular as new technologies become available, especially as some are customized for plants (*107*).

Concluding remarks. In the initial annotation of the human genome (*7*, *8*), alternative splicing was proposed as a method by which protein diversity could be generated from the surprisingly small number of genes that were identified. The idea that there is extensive alternative splicing in human genes has been supported by analyses of EST data (*108*–*112*). Alternative splicing is often associated with the exon recognition model (*113*) of pre-mRNA splicing. Exon recognition is facilitated by exonic splicing enhancers—short, degenerate sequences located in the exons that are recognized by a multitude of RNA binding factors (*114*, *115*). Because it is the exons that are recognized by the splicing machinery, the intron sequence content is less critical, and transposon insertions into the intron are more readily tolerated. Thus, the preponderance of large transposons-filled introns in the human genome is consistent with extensive alternative splicing.

The presence of relatively few transposons inside plant introns suggests that exon recognition is not a common process for plant genes. Indeed, exonic splicing enhancers have yet to be identified in plants (*116*). The corollary is that there should be relatively little alternative splicing in plant genes. Analysis of the EST data confirms that *A. thaliana* has substantially less alternative splicing than vertebrates or invertebrates (*117*). However,

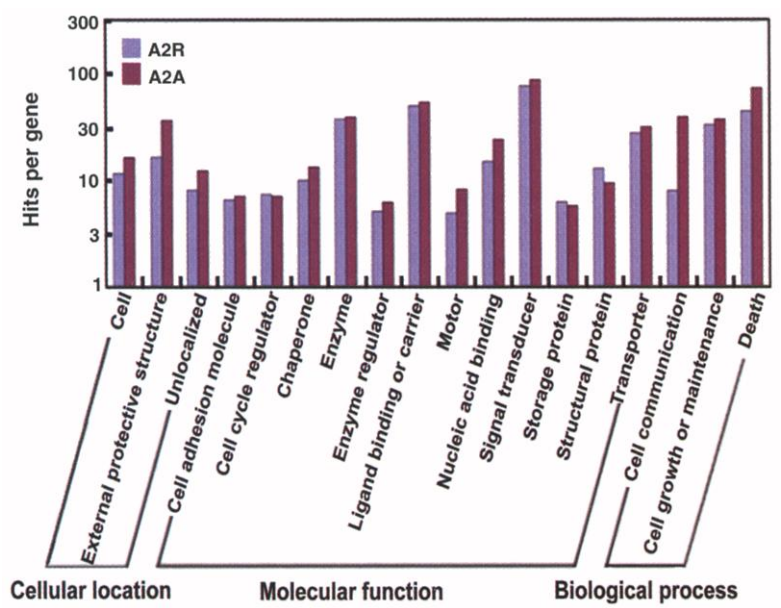

Fig. 13. Distributions in the number of hits per gene, sorted according to Gene Ontology Consortium, for *Arabidopsis*-to-rice and *Arabidopsis*-to-*Arabidopsis* comparisons. This figure shows only the 36.3% of predicted *A. thaliana* genes that are classified.

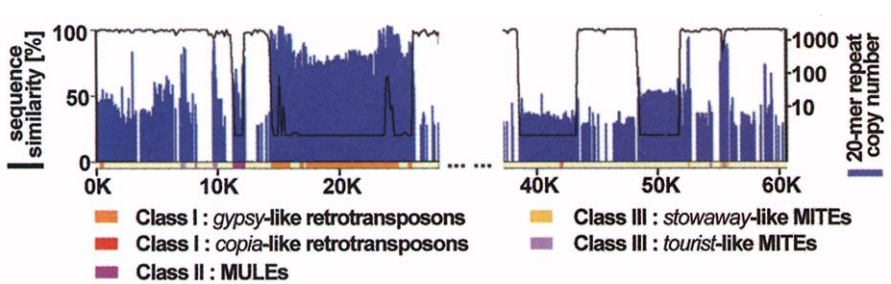

Fig. 14. Comparison of *GLA* (*indica*) and *Nipponbare* (*japonica*) BAC sequences (GenBank accession number AL442110 and AL606449, respectively). Exact 20-mer repeats are indicated by blue histogram bars, with bar heights proportional to copy number in *93-11* (*indica*). Sequence similarity is almost 100%, or unalignable and set to 20% by BLAST. Every unalignable region coincides with a cluster of MDRs, but RepeatMasker fails to identify a BDR in half of these regions.

protein diversity must be generated for the organism to evolve. Our analysis has demonstrated extensive gene duplications in rice and *A. thaliana*, which are highly correlated with each other when decomposed by functional classification. The conclusion is that protein diversity in plants is generated primarily through gene duplications, whereas in vertebrates, it is generated through gene duplications and alternative splicing. This would explain why rice has so many genes. However, as a method of generating protein diversity, gene duplications come at the cost of an increase in transcriptional noise (*118*). Perhaps, at some level of complexity, alternative splicing becomes preferred.

Looking to the future, we intend to improve our draft sequence by adding more reads from large-insert clones, filling any gaps that are likely to contain genes, and integrating the resultant sequence with existing physical and genetic maps. The large-insert clones are necessary to correctly assemble across the large repeat clusters that are sprinkled throughout the rice genome. Until then, the BAC-end sequences (*119*) may not be useful because they are too large to bridge adjacent contigs, and instead skip intervening contigs, resulting in a morass of interleaving scaffolds. One should also be wary of large-scale differences between *indica* and *japonica*. In any event, the final assembly will be made freely available to the research community. We will then apply the experiences gained from the rice genome project to other agriculturally important crops, including *Z. mays* (maize) and *T. aestivum* (wheat).

References and Notes

1. T. Sasaki, B. Burr, *Curr. Opin. Plant Biol.* **3**, 138 (2000).
2. N. A. Eckardt, *Plant Cell* **12**, 2011 (2000).
3. K. Arumuganathan, E. D. Earle, *Plant Mol. Biol. Rep.* **9**, 208 (1991).
4. The *Arabidopsis* Genome Initiative, *Nature* **408**, 796 (2000).
5. B. Martienssen, W. R. McCombie, *Cell* **10**, 571 (2001).
6. M. Bevan *et al.*, *Curr. Opin. Plant Biol.* **4**, 105 (2001).
7. International Human Genome Sequencing Consortium, *Nature* **409**, 860 (2001).
8. J. C. Venter *et al.*, *Science* **291**, 1304 (2001).
9. Q. Tao *et al.*, *Cell Res.* **4**, 127 (1994).
10. Y. Umehara, A. Miyazaki, H. Tanoue, *Mol. Breed.* **1**, 79 (1995).
11. M. Bevan, G. Murphy, *Trends Genet.* **15**, 211 (1999).
12. G. L. Wang *et al.*, *Plant J.* **7**, 525 (1995).
13. M. D. Gale, K. M. Devos, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 1971 (1998).
14. J. Messing, V. Llaca, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 2017 (1998).
15. S. Goff, *Curr. Opin. Plant Biol.* **2**, 86 (1999).
16. Major Web sites for rice genome data: http://rgp.dna.affrc.go.jp; http://www.genome.clemson.edu; http://ars-genome.cornell.edu; http://www.tigr.org/tdb/e2k1/osa1/BACmapping/description.shtml.
17. R. J. Davenport, *Science* **291**, 807 (2001).
18. D. Dickson, D. Cyranoski, *Nature* **409**, 551 (2001).
19. Z. Y. Dai, B. H. Zhao, X. J. Liu (in Chinese), *Jiangsu Agric. Sci.* **4**, 13 (1997).
20. L. P. Yuan (in Chinese), *Hybrid Rice* **1**, 1 (1997).
21. J. Yu *et al.*, *Chin. Sci. Bull.* **46**, 1937 (2001).
22. J. Wang *et al.*, *Genome Res.*, in press. The software can be obtained by e-mailing the authors at reps@genomics.org.cn.
23. J. L. Bennetzen, *Plant Cell* **12**, 1021 (2000).
24. A. Kumar, J. L. Bennetzen, *Annu. Rev. Genet.* **33**, 479 (1999).
25. M. D. Adams *et al.*, *Science* **287**, 2185 (2000).
26. B. Ewing, L. Hillier, M. C. Wendl, P. Green, *Genome Res.* **8**, 175 (1998).
27. B. Ewing, P. Green, *Genome Res.* **8**, 186 (1998).
28. For the plasmid shotgun libraries, a DNA isolation protocol was modified from Sambrook and Russell (*29*). Fresh leaves at the seeding stage were ground in liquid nitrogen before complete lysis (*30*). Purified high–molecular weight genomic DNA was sonicated and sized on agarose gels, selecting for fragments of size 1.5 to 3.0 kb. QIAEX Gel Extraction Kit (QIAGEN) was used to purify DNA from the gel slices. Genomic fragments were ligated to Sma I–linearized pUC18 plasmids and transformed into DH10B-competent cells by electroporation.
29. J. Sambrook, J. D. Russell, *Molecular Cloning*, (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, ed. 3, 2001).
30. S. Hatano, J. Yamaguchi, A. Hirai, *Plant Sci.* **83**, 55 (1992).
31. Single colonies were grown in 96-deep-well plates, and plasmid DNA was prepared by alkaline lysis (*32*). Quality of DNA and insert sizes were examined by agarose gel electrophoresis. Purified plasmid DNA (200 ng; Amersham Pharmacia Biotech, Beijing) was used for the sequencing reactions. DNA sequencing was done with MegaBACE 1000 capillary sequencers (Amersham Pharmacia Biotech, Beijing). Machine parameters were adjusted for high output (10 to 11 runs a day on average).
32. H. C. Birnboim, *Methods Enzymol.* **100**, 243 (1983).
33. N. Jiang, S. R. Wessler, *Plant Cell* **13**, 2553 (2001).
34. P. Green, http://www.phrap.org.
35. E. S. Lander, M. S. Waterman, *Genomics* **2**, 231 (1988).
36. S. F. Altschul, W. Gish, *Methods Enzymol.* **266**, 460 (1996).
37. Sources for STS, STR, restriction fragment length polymorphism sequences: http://ars-genome.cornell.edu/; http://www.ncbi.nlm.nih.gov/; http://rgp.dna.affrc.go.jp/publicdata/geneticmap2000/index.html.
38. A. F. Smit, P. Green, http://ftp.genome.washington.edu/RM/RepeatMasker.html.
39. S. P. Kennedy, W. V. Ng, S. L. Salzberg, L. Hood, S. DasSarma, *Genome Res.* **11**, 1641 (2001).
40. E. Chargaff, *Experientia* **6**, 201 (1950).
41. R. Rolfe, M. Meselson, *Proc. Natl. Acad. Sci. U.S.A.* **45**, 1039 (1959).
42. N. Sueoka, J. Marmur, P. Doty, *Nature* **183**, 1427 (1959).
43. D. R. Forsdyke, J. R. Mortimer, *Gene* **261**, 127 (2000).
44. G. Bernardi, *Gene* **259**, 31 (2000).
45. A. Eyre-Walker, L. D. Hurst, *Nature Rev. Genet.* **2**, 540 (2001).
46. C. Gautier, *Curr. Opin. Genet. Dev.* **10**, 656 (2000).
47. S. Karlin, A. M. Campbell, J. Mrazek, *Annu. Rev. Genet.* **32**, 185 (1998)
48. N. Carels, G. Bernardi, *Genetics* **154**, 1819 (2000).
49. J. Filipski, J. P. Thiery, G. Bernardi, *J. Mol. Biol.* **80**, 177 (1973).
50. G. Bernardi *et al.*, *Science* **228**, 953 (1985).
51. G. K. S Wong, D. A. Passey, Y. Z. Huang, Z. Yang, J. Yu, *Genome Res.* **10**, 1672 (2000).
52. G. K. S. Wong *et al.*, *Genome Res.*, in press.
53. G. K. S. Wong, D. A. Passey, J. Yu, *Genome Res.* **11**, 1672 (2001).
54. C. Feuillet, B. Keller, *Proc. Natl. Acad. Sci. U.S.A.* **96**, 8265 (1999).
55. I. Ashikawa, *Plant J.* **26**, 617 (2001).
56. F. Larsen, G. Gundersen, L. Lopez, H. Prydz, *Genomics* **13**, 1095 (1992).
57. J. Messing, *Trends Genet.* **6**, 196 (2001).
58. R. Sánchez-Fernández, P. A. Rea, T. G. E. Davies, J. O. D. Coleman, *Trends Plant Sci.* **6**, 348 (2001).
59. J. Jurka, *Trends Genet.* **16**, 418 (2000); http://www.girinst.org/index.html.
60. L. Mao *et al.*, *Genome Res.* **10**, 982 (2000).
61. K. Turcotte, S. Srinivasan, T. Bureau, *Plant J.* **25**, 169 (2001).
62. S. Temnykh *et al.*, *Genome Res.* **11**, 1441 (2001).
63. R. I. Richards, G. R. Sutherland, *Cell* **70**, 709 (1992).
64. ———, *Nature Genet.* **6**, 114 (1994).
65. C. Schlötterer, D. Tautz, *Nucleic Acids Res.* **20**, 211 (1992).
66. D. Tautz, M. Trick, G. Dover, *Nature* **322**, 652 (1986).
67. S. A. Surzycki, W. R. Belknap, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 245 (2000).
68. R. Tarchini, P. Biddle, R. Wineland, S. Tingey, A. Rafalski, *Plant Cell* **12**, 381 (2000).
69. X. Zhang *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **98**, 12572 (2001).
70. A. F. Smit, *Curr. Opin. Genet. Dev.* **6**, 743 (1996).
71. T. E. Bureau, P. C. Ronald, S. R. Wessler, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 8524 (1996).
72. A. Nekrutenko, W. H. Li, *Trends Genet.* **17**, 619 (2001).
73. M. Long, *Curr. Opin. Genet. Dev.* **11**, 673 (2001).
74. F. Jacob, *Science* **196**, 1141 (1977).
75. A. A. Salamov, V. V. Solovyev, *Genome Res.* **10**, 516 (2000); http://www.softberry.com/berry.phtml?topic=gfind.
76. A. V. Lukashin, M. Borodovsky, *Nucleic Acids Res.* **26**, 1107 (1998). http://opal.biology.gatech.edu/GeneMark/eukhmm.cgi?org=O.sativa.
77. C. Burge, S. Karlin, *J. Mol. Biol.* **268**, 78 (1997); http://genes.mit.edu/GENSCAN.html.
78. A. L. Delcher, D. Harmon, S. Kasif, O. White, S. L. Salzberg, *Nucleic Acids Res.* **27**, 4636 (1999); http://www.tigr.org/softlab/glimmerm.
79. K. Sakata *et al.*, *Abstracts of 4th Annual Conference on Computational Genomics* (2000), p. 31; http://rgp.dna.affrc.go.jp/RiceHMM.
80. S. Rogic, A. K. Mackworth, F. B. Ouellette, *Genome Res.* **11**, 817 (2001).
81. P. Breyne, M. Zabeau, *Curr. Opin. Plant Biol.* **4**, 42 (2001).
82. R. Apweiler *et al.*, *Nucleic Acids Res.* **29**, 44 (2001).
83. The Gene Ontology Consortium, *Nature Genet.* **25**, 25 (2000).
84. M. D. Gale, K. M. Devos, *Science* **282**, 656 (1998).
85. R. Schmidt, *Curr. Opin. Plant Biol.* **3**, 97 (2000).
86. K. M. Devos, J. Beales, Y. Nagamura, T. Sasaki, *Genome Res.* **9**, 825 (1999).
87. A. van Dodeweerd *et al.*, *Genome* **42**, 887 (1999).
88. K. Mayer *et al.*, *Genome Res.* **11**, 1167 (2001).
89. H. Liu, R. Sachidanandam, L. Stein, *Genome Res.* **11**, 2020 (2001).
90. V. Brendel, S. Kurtz, V. Walbot, *Genome Biol.* **3**, reviews 1005.1 (2002).
91. M. Devan *et al.*, *Curr. Opin. Plant Biol.* **4**, 105 (2001).
92. W. Fitch, *Syst. Zool.* **19**, 99 (1970).
93. *Arabidopsis* genome annotations: http://www.arabidopsis.org.
94. E. Gasteiger, E. Jung, A. Bairoch, *Curr. Iss. Mol. Biol.* **3**, 47 (2001); http://www.expasy.com.
95. For each protein query, we created an array with one element for each amino acid position. Blast_hits() recorded the number of times that each position was covered by a TblastN hit. Each hit had associated with it a score for the percentage of identically matched amino acids. AA_identity() recorded the maximum and minimum score at each position, across all TblastN hits. "Extent of hit," quoted as a percentage of the protein length, is the number of nonzero elements in Blast_hits(). "AA identity" and "hits per gene" are the median values of AA_identity() and Blast_hits(), computed over positions with one or more hits. We used the median, instead of the mean, to minimize the likelihood of counting a highly duplicated domain when the entire protein is not duplicated.
96. S. A. Islam, J. Luo, M. J. Sternberg, *Protein Eng.* **8**, 513 (1995).
97. R. Sowdhamini, S. D. Rufino, T. L. Blundell, *Fold Des.* **1**, 209 (1996).
98. S. J. Wheelan, A. Marchler-Bauer, S. H. Bryant. *Bioinformatics* **16**, 613 (2000).
99. S. Ge, T. Sang, B. R. Lu, D. Y. Hong, *Proc. Natl. Acad. Sci. U.S.A.* **96**, 14400 (1999).
100. L. H. Rieseberg, *Annu. Rev. Ecol. Syst.* **28**, 359 (1997).
101. J. Masterson, *Science* **264**, 421 (1994).
102. L. Comai, *Plant Mol. Biol.* **43**, 387 (2000).
103. C. M. Vicient, M. J. Jaaskelainen, R. Kalendar, A. H. Schulman, *Plant Physiol.* **125**, 1283 (2001).
104. M. I. Tenaillon *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **98**, 9161 (2001).
105. L. Kruglyak, *Nature Genet.* **17**, 21 (1997).

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).
→ 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.