

The map-based sequence of the rice genome

International Rice Genome Sequencing Project*

Rice, one of the world's most important food plants, has important syntenic relationships with the other cereal species and is a model plant for the grasses. Here we present a map-based, finished quality sequence that covers 95% of the 389 Mb genome, including virtually all of the euchromatin and two complete centromeres. A total of 37,544 non-transposable-element-related protein-coding genes were identified, of which 71% had a putative homologue in *Arabidopsis*. In a reciprocal analysis, 90% of the *Arabidopsis* proteins had a putative homologue in the predicted rice proteome. Twenty-nine per cent of the 37,544 predicted genes appear in clustered gene families. The number and classes of transposable elements found in the rice genome are consistent with the expansion of syntenic regions in the maize and sorghum genomes. We find evidence for widespread and recurrent gene transfer from the organelles to the nuclear chromosomes. The map-based sequence has proven useful for the identification of genes underlying agronomic traits. The additional single-nucleotide polymorphisms and simple sequence repeats identified in our study should accelerate improvements in rice production.

Rice (*Oryza sativa* L.) is the most important food crop in the world and feeds over half of the global population. As the first step in a systematic and complete functional characterization of the rice genome, the International Rice Genome Sequencing Project (IRGSP) has generated and analysed a highly accurate finished sequence of the rice genome that is anchored to the genetic map. Our analysis has revealed several salient features of the rice genome:

- We provide evidence for a genome size of 389 Mb. This size estimation is ~260 Mb larger than the fully sequenced dicot plant model *Arabidopsis thaliana*. We generated 370 Mb of finished sequence, representing 95% coverage of the genome and virtually all of the euchromatic regions.
- A total of 37,544 non-transposable-element-related protein-coding sequences were detected, compared with ~28,000–29,000 in *Arabidopsis*, with a lower gene density of one gene per 9.9 kb in rice. A total of 2,859 genes seem to be unique to rice and the other cereals, some of which might differentiate monocot and dicot lineages.
- Gene knockouts are useful tools for determining gene function and relating genes to phenotypes. We identified 11,487 *Tos17* retrotransposon insertion sites, of which 3,243 are in genes.
- Between 0.38 and 0.43% of the nuclear genome contains organellar DNA fragments, representing repeated and ongoing transfer of organellar DNA to the nuclear genome.
- The transposon content of rice is at least 35% and is populated by representatives from all known transposon superfamilies.
- We have identified 80,127 polymorphic sites that distinguish between two cultivated rice subspecies, *japonica* and *indica*, resulting in a high-resolution genetic map for rice. Single-nucleotide polymorphism (SNP) frequency varies from 0.53 to 0.78%, which is 20 times the frequency observed between the Columbia and Landsberg *erecta* ecotypes of *Arabidopsis*.
- A comparison between the IRGSP genome sequence and the

6.3 × *indica* and 6 × *japonica* whole-genome shotgun sequence assemblies revealed that the draft sequences provided coverage of 69% by *indica* and 78% by *japonica* relative to the map-based sequence.

Rice has played a central role in human nutrition and culture for the past 10,000 years. It has been estimated that world rice production must increase by 30% over the next 20 years to meet projected demands from population increase and economic development¹. Rice grown on the most productive irrigated land has achieved nearly maximum production with current strains¹. Environmental degradation, including pollution, increase in night time temperature due to global warming², reductions in suitable arable land, water, labour and energy-dependent fertilizer provide additional constraints. These factors make steps to maximize rice productivity particularly important. Increasing yield potential and yield stability will come from a combination of biotechnology and improved conventional breeding. Both will be dependent on a high-quality rice genome sequence.

Rice benefits from having the smallest genome of the major cereals, dense genetic maps and relative ease of genetic transformation³. The discovery of extensive genome colinearity among the Poaceae⁴ has established rice as the model organism for the cereal grasses. These properties, along with the finished sequence and other tools under development, set the stage for a complete functional characterization of the rice genome.

The International Rice Genome Sequencing Project

The IRGSP, formally established in 1998, pooled the resources of sequencing groups in ten nations to obtain a complete finished quality sequence of the rice genome (*Oryza sativa* L. ssp. *japonica* cv. Nipponbare). Finished quality sequence is defined as containing less than one error in 10,000 nucleotides, having resolved ambiguities, and having made all state-of-the-art attempts to close gaps. The IRGSP released a high-quality map-based draft sequence in

*Lists of participants and affiliations appear at the end of the paper

December 2002. Three completely sequenced chromosomes have been published^{5,7}, as well as two completely sequenced centromeres^{8,10}. As the IRGSP subscribed to an immediate-release policy, high-quality map-based sequence has been public for some time. This has permitted rice geneticists to identify several genes underlying traits, and revealed very large and previously unknown segmental duplications that comprise 60% of the genome^{11,13}. The public sequence has also revealed new details about the synteny relationships and gene mobility between rice, maize and sorghum^{13,15}.

Physical maps, sequencing and coverage

The IRGSP sequenced the genome of a single inbred cultivar, *Oryza sativa* ssp. *japonica* cv. Nipponbare, and adopted a hierarchical clone-by-clone method using bacterial and plasmid artificial chromosome clones (BACs and PACs, respectively). This strategy used a high-density genetic map¹⁶, expressed-sequence tags (ESTs)¹⁷, yeast artificial chromosome (YAC)- and BAC-based physical maps^{18,20}, BAC-end sequences²¹ and two draft sequences^{22,23}. A total of 3,401 BAC/PAC clones (Table 1) were sequenced to approximately tenfold sequence coverage, assembled, ordered and finished to a sequence quality of less than one error per 10,000 bases. A majority of physical gaps in the BAC/PAC tiling path were bridged using a variety of substrates, including PCR fragments, 10-kb plasmids and 40-kb fosmid clones. A total of 62 unsequenced physical gaps, including nine centromere and 17 telomere gaps, remain on the 12 chromosomes (Table 2). Chromosome arm and telomere gaps were measured, and the nine centromere gaps were estimated on the basis of CentO satellite DNA content. The remaining gaps are estimated to total 18.1 Mb.

Nineteen-seven percent of the BAC/PACs and gap sequences (3,360) have been submitted as finished quality in the PLN division of GenBank/DDBJ/EMBL. These and the remaining draft-sequenced clones were used to construct pseudomolecules representing the 12 chromosomes of rice (Fig. 1). The total nucleotide sequence of the 12 pseudomolecules is 370,733,456 bp, with an N-average continuous sequence length of 6.9 Mb (see Table 1 for a definition of N-average length). Sequence quality was assessed by comparing 1.2 Mb of overlapping sequence produced by different laboratories. The overall accuracy was calculated as 99.99% (Supplementar Table 2). The statistics of sequenced PAC/BAC clones and pseudomolecules for each chromosome are shown in Table 1.

The genome size of rice (*O. sativa* ssp. *japonica* cv. Nipponbare) was reported to have a haploid nuclear DNA content of 394 Mb on the basis of flow cytometry²⁴, and 403 Mb on the basis of lengths of anchored BAC contigs and estimates of gap sizes²⁰. Table 2 shows the calculated size for each chromosome and the estimated coverage. Adding the estimated length of the gaps to the sum of the non-overlapping sequence, the total length of the rice nuclear genome was calculated to be 388.8 Mb. Therefore, the pseudomolecules are expected to cover 95.3% of the entire genome and an estimated 98.9% of the euchromatin. An independent measure of genome coverage represented by the pseudomolecules was obtained by searching for unique EST markers¹⁹; of 8,440 ESTs, 8,391 (99.4%) were identified in the pseudomolecules.

Centromere location

Typical eukaryotic centromeres contain repetitive sequences, including satellite DNA at the centre and retrotransposons and transposons in the flanking regions. All rice centromeres contain the highly repetitive 155–165 bp CentO satellite DNA, together with centromere-specific retrotransposons^{25,26}. The CentO satellites are located within the functional domain of the rice centromere^{10,26}. Complete sequencing of the centromeres of rice chromosomes 4 and 8 revealed that they consist of 59 kb and 69 kb of clustered CentO repeats (respectively)^{8,10}, tandemly arranged head-to-tail within the clusters. Numerous retrotransposons, including the centromere-specific

RIRE7, are found between and around the CentO repeats. CentO clusters show differences in length and orientation for the two centromeres.

BLASTN analysis of the pseudomolecules indicated that about 0.9 Mb of CentO repeats (corresponding to more than 5,800 copies of the satellite) were sequenced and found to be associated with centromere-specific retroelements. Locations of all CentO sequences correspond to genetically identified centromere regions (Supplementar Table 3). Our pseudomolecules cover the centromere regions on chromosomes 4, 5 and 8, and portions of the centromeres on the remaining chromosomes (Fig. 1).

Gene content, expression and distribution

We masked the pseudomolecules for repetitive sequences and used the *ab initio* gene finder FGENESH to identify only non-transposable-element-related genes. A total of 37,544 non-transposable-element protein-coding sequences were predicted, resulting in a density of one gene per 9.9 kb (Supplementar Tables 4 and 5). As the ability to identify unannotated and transposable-element-related genes improves, the true protein-coding gene number in rice will doubtless be revised.

Full-length complementary DNA sequences are available for rice²⁷, and provide a powerful resource for improving gene model structure derived from *ab initio* gene finders²⁸. Of the 37,544 non-transposable-element-related FGENESH models, 17,016 could be supported by a total of 25,636 full-length cDNAs (Supplementar Table 6).

A total of 22,840 (61%) genes had a high identity match with a rice EST or full-length cDNA. On average, about 10.7 EST sequences were present for each expressed rice gene. A total of 2,927 genes aligned well with ESTs from other cereal species, and 330 of these genes matched only with a non-rice cereal EST (Supplementar Fig. 1). Except for the short arms of chromosomes 4, 9 and 10, which are known to be highly heterochromatic, the density of expressed genes is greater on the distal portions of the chromosome arms compared with the regions around the centromeres (Supplementar Fig. 2).

A total of 19,675 proteins had matches with entries in the Swiss-Prot database; of these, 4,500 had no expression support. Domain searches revealed a minimum of one motif or domain present in 63% of the predicted proteins, with a total of 3,328 different domains present in the predicted rice proteome. The five most abundant domains were associated with protein kinases (Supplementar Table 7). Fifteen-one per cent of the predicted proteins could be associated with a biological process (Supplementar Fig. 3a), with metabolism (29.1%) and cellular physiological processes (11.9%) representing the two most abundant classes.

Approximately 71% (26,837) of the predicted rice proteins have a homologue in the *Arabidopsis* proteome (Supplementar Fig. 4). In a reciprocal search, 89.8% (26,004) of the proteins from the *Arabidopsis* genome have a homologue in the rice proteome. Of the 23,170 rice genes with rice EST, cereal EST, or full-length cDNA support, 20,311 (88%) have a homologue in *Arabidopsis*. Fewer putative homologues were found in other model species: 38.1% in *Drosophila*, 40.8% in human, 36.5% in *Caenorhabditis elegans*, 30.2% in yeast, 17.6% in *Synechocystis* and 10.2% in *Escherichia coli*.

There are profound differences in plant architecture and biochemistry between monocotyledonous and dicotyledonous angiosperms. Only 2,859 rice genes with evidence of transcription lack homologues in the *Arabidopsis* genome. We investigated these to learn what functions they encoded. The vast majority had no matches, or most closely matched unknown or hypothetical proteins. The grasses have a class of seed storage proteins called prolamins that is not found in dicots. There are also families of hormone response proteins and defence proteins, such as proteinase inhibitors, chitinases, pathogenesis-related proteins and seed allergens, many of which are tandemly repeated (Supplementar Table 8). Nevertheless, with a large number of proteins of unknown function, the most interesting

differences between the genome content of these two groups of angiosperms remain to be discovered.

Tos17 is an endogenous *cop*ia-like retrotransposon in rice that is inactive under normal growth conditions. In tissue culture, it becomes activated, transposes and is stably inherited when the plant is regenerated²⁹. There are only two copies of *Tos17* in the rice cultivar Nipponbare. These features, together with its preferential insertion into gene-rich regions, make *Tos17* uniquely suitable for the functional analysis of rice genes by gene disruption. About 50,000 *Tos17*-insertion lines carrying 500,000 insertions have been produced³⁰. A total of 11,487 target loci were mapped on the 12 pseudomolecules (Supplementar Fig. 5), with at least one insertion detected in 3,243 genes. The density of *Tos17* insertions is higher in euchromatic regions of the genome³⁰, in contrast to the distribution of high-copy retrotransposons, which are more frequently found in pericentromeric regions. A similar target site preference has been reported for T-DNA insertions in *Arabidopsis*³¹.

Tandem gene families

One surprising outcome of the *Arabidopsis* genome analysis was the large percentage (17%) of genes arranged in tandem repeats³². When performing a similar analysis with rice, the percentage was comparable (14%). However, manual curation on rice chromosome 10 showed one gene family encoding a glucanase-rich protein with 27 copies and one encoding a TRAF/BTB domain protein with 48 copies³³. These tandemly repeated families are interrupted with other genes and are not included in strictly defined tandem repeats. We therefore screened for all tandemly arranged genes in 5-Mb intervals. Using these criteria, 29% of the genes (10,837) are amplified at least once in tandem, and 153 rice gene arrays contained 10–134 members (Supplementar Fig. 6). Sixty per cent of the tandem arrays with over 27 members, and 33% of all the arrays with over 10 members, contain protein kinase domains (Supplementar Table 9).

Non-coding RNA genes

The nucleolar organiser, consisting of 17S–5.8S–25S ribosomal DNA coding units, is found at the telomeric end of the short arm of chromosome 9 (ref. 34) in *O. sativa* ssp. *japonica*, and is estimated to comprise 7 Mb (ref. 35). A second 17S–5.8S–25S rDNA locus is found at the end of the short arm of chromosome 10 in *O. sativa* ssp.

*indica*³⁴. A single 5S cluster is present on the short arm of chromosome 11 in the vicinity of the centromere³⁶, and encompasses 0.25 Mb.

A total of 763 transfer RNA genes, including 14 tRNA pseudogenes were detected in the 12 pseudomolecules. In comparison, a total of 611 tRNA genes were detected in *Arabidopsis*³². Supplementar Fig. 7 shows the distribution of these tRNA genes in each chromosome. Chromosome 4 has a single tRNA cluster⁶, and chromosome 10 has two large clusters derived from inserted chloroplast DNA⁷. Except for regions of intermediate density on chromosomes 1, 2, 8 and 12, there seem to be no other large clusters.

MicroRNAs (miRNAs), a class of eukaryotic non-coding RNAs, are believed to regulate gene expression by interacting with the target messenger RNA³⁷. miRNAs have been predicted from *Arabidopsis*³⁸ and rice³⁹, and we mapped 158 miRNAs onto the rice pseudomolecules (Supplementar Table 10). Among other non-coding RNAs, we identified 215 small nucleolar RNA (snoRNA) and 93 spliceosomal RNA genes, both showing biased chromosomal distributions, in the rice genome (Supplementar Table 11).

Organelle insertions in the nuclear genome

Mitochondria and chloroplasts originated from alpha-proteobacteria and cyanobacteria endosymbionts. A continuous transfer of organellar DNA to the nucleus has resulted in the presence of chloroplast and mitochondrial DNA inserted in the nuclear chromosomes. Although the endosymbionts probably contained genomes of several Mb at the time they were internalised, the organellar genomes diminished so that the present size of the mitochondrial genome is less than 600 kb, and that of the chloroplast is only 150 kb. Homolog searches detected 421–453 chloroplast insertions and 909–1,191 mitochondrial insertions, depending upon the stringency adopted (Supplementar Fig. 8 and Supplementar Table 12). Thus, chloroplast and mitochondrial insertions contribute 0.20–0.24% and 0.18–0.19% of the nuclear genome of rice, respectively, and correspond to 5.3 chloroplast and 1.3 mitochondrial genome equivalents. The distribution of chloroplast and mitochondrial insertions over the 12 chromosomes indicates that mitochondrial and chloroplast transfers occurred independently. Two chromosomes harbour more insertions than the others (Supplementar Fig. 8 and Supplementar Table 12), with chromosome 12 containing nearly 1% mitochondrial DNA and chromosome 10 containing approximately 0.8% chlor-

Table 1 | Classification and distribution of sequenced PAC and BAC clones* on the 12 rice chromosomes

Chr	Sequencing laboratory†	PAC	BAC	OSJNBa/b	OJ	OSJNO	Others‡	Total§	Pseudomolecule (bp)	N-average length (bp)	Accession no.
1	RGP, KRGRP	251	77	42	23	4	0	397	43,260,640	9,688,259	AP008207
2	RGP, JIC	117	16	80	142	4	0	359	35,954,074	7,793,366	AP008208
3	ACWW, TIGR	1	8	263	47	1	10	330	36,189,985	5,196,992	AP008209
4	NCGR	2	7	275	7	0	0	291	35,489,479	1,427,419	AP008210
5	ASPGC	67	11	113	87	0	0	278	29,733,216	3,086,418	AP008211
6	RGP	169	20	78	14	0	0	281	30,731,386	8,669,608	AP008212
7	RGP	102	19	68	97	0	0	286	29,643,843	14,923,781	AP008213
8	RGP	113	23	56	83	2	0	277	28,434,680	14,872,702	AP008214
9	RGP, KRGRP, BIOTEC, BRIGI	72	24	72	50	5	0	223	22,692,709	5,219,517	AP008215
10	ACWW, TIGR, PGIR	1	5	172	6	0	21	205	22,683,701	2,124,647	AP008216
11	ACWW, TIGR, IIRGS, PGIR, Genoscope	10	6	236	3	2	1	258	28,357,783	1,087,274	AP008217
12	Genoscope	2	6	179	79	0	2	268	27,561,960	7,600,514	AP008218
	Total	907	222	1634	638	18	34	3453	370,733,456	6,928,182	

Chr, chromosome.

*PAC, Rice Genome Research Program PAC; BAC, Rice Genome Research Program BAC; OSJNBa/b, Clemson University Genomics Institute BAC; OJ, Monsanto BAC; OSJNO, Arizona Genomics Institute fosmid (<http://www.genome.arizona.edu/orders/direct.html?library=OSJNOa>); Others, artificial gap-filling clones designated as OSJNA and OJA.

†ACWW (Arizona Genomics Institute, Cold Spring Harbor Laboratory, Washington University Genome Sequencing Center, University of Wisconsin) Rice Genome Sequencing Consortium; ASPGC, Academia Sinica Plant Genome Center; BIOTEC, National Center for Genetic Engineering and Biotechnology; BRIGI, Brazilian Rice Genome Initiative; IIRGS, Indian Initiative for Rice Genome Sequencing; JIC, John Innes Centre; KRGRP, Korea Rice Genome Research Program; NCGR, National Center for Gene Research; PGIR, Plant Genome Initiative at Rutgers; RGP, Rice Genome Research Program; TIGR, The Institute for Genomic Research.

‡Constructs derived by joining (mostly from the clone gap regions) sequence from PCR fragments, Monsanto or Syngenta sequences and the neighbouring clone sequences.

§A total of 2,494 BAC and 907 PAC clones were used for draft and finished sequencing. Monsanto draft-sequenced BACs underlie 638 finished clones. The Syngenta draft sequence contributed to the assemblies of 140 IRGSP clone sequences. Thirty-four sequence submissions are artificial constructs derived by joining a regional sequence (mostly from the clone gap regions) from PCR fragments, Monsanto or Syngenta sequences with the neighbouring clone sequences. This also includes 93 clones submitted as phase 1 or phase 2 to the HTG section of GenBank.

||N-average length: the average length of a contiguous segment (without sequence or physical gaps) containing a randomly chosen nucleotide.

oplast DNA. It is clear that several successive transfer events have occurred, as insertions of less than 10 kb have heterogeneous identities. The longest insertions, however, systematically show >98.5% identity to organellar DNA (Supplementar Table 13), indicating recent insertions for both chloroplast and mitochondrial genomes.

Transposable elements

The rice genome is populated by representatives from all known transposon superfamilies, including elements that cannot be easily classified into either class I or II (ref. 40). Previous estimates of the transposon content in the rice genome range from 10 to 25% (refs 21, 40). However, the increased availability of transposon queries and the use of profile hidden Markov models allow the identification of more divergent elements⁴¹ and indicate that the transposon content of the *O. sativa* ssp. *japonica* genome is at least 35% (Table 3). Chromosomes 8 and 12 have the highest transposon content (38.0% and 38.3%, respectively), and chromosomes 1 (31.0%), 2 (29.8%) and 3 (29.0%) have the lowest proportion of transposons. Conversely, elements belonging to the IS5/*Tourist* and IS630/*Tc1/mariner* superfamilies, which are generally correlated with gene density, are prevalent on the first three chromosomes and least frequent on chromosomes 4 and 12.

Class II elements, characterized by terminal inverted-repeats and including the *hAT*, *CACTA*, *IS256/Mutator*, *IS5/Tourist*, and *IS630/Tc1/mariner* superfamilies, outnumber class I elements, which include long terminal-repeat (LTR) retrotransposons (*T1/copia*, *T3/gypsy* and *TRIM*) and non-LTR retrotransposons (LINEs and SINEs, or long- and short-interspersed nucleotide elements, respectively), by more than twofold (Table 3). However, the nucleotide contribution of class I is greater than that of class II, due mostly to the large size of LTR retrotransposons and the small size of IS5/*Tourist* and IS630/*Tc1/mariner* elements. The inverse is the case for maize, for which class I elements outnumber class II elements⁴². Given their larger sizes, differential amplification of LTR elements in maize compared with rice is consistent with the genomic expansion found between orthologous regions of rice and maize^{15,33}.

Most class I elements are concentrated in gene-poor, heterochromatic regions such as the centromeric and pericentromeric regions (Supplementar Table 14). In contrast, members of some transposon superfamilies, including IS5/*Tourist*, IS630/*Tc1/mariner* and LINEs, have a significant positive correlation with both recombination rate and gene density. There is an effect of average element length associated with these patterns: short elements generally show a positive correlation with recombination rate and gene density, and are under-represented in the centromere regions, whereas larger elements have higher centromeric and pericentromeric abundance.

Intraspecific sequence polymorphism

Map-based cloning to identify genes that are associated with agronomic traits is dependent on having a high frequency of polymorphic markers to order recombination events. In rice, most of the segregating populations are generated from crosses between the two major subspecies of cultivated rice, *Oryza sativa* ssp. *japonica* and *O. sativa* ssp. *indica*. Although several studies on the polymorphisms detected between *japonica* and *indica* subspecies have been reported^{6,43,44}, the analysis reported here uses an approach that ensures comparison of orthologous sequences. *O. sativa* ssp. *indica* cv. Kasalath and *O. sativa* ssp. *japonica* cv. Nipponbare are the parents of the most densely mapped rice population¹⁶. BAC-end sequences were obtained from a Kasalath BAC library of 47,194 clones. Only high quality, single-copy sequences were mapped to the Nipponbare pseudomolecules, and only paired inverted sequences that mapped within 200 kb were considered. A total of 26,632 paired Kasalath BAC-end sequences were mapped to the 12 rice pseudomolecules (Supplementar Table 15). Kasalath BAC clones spanned 308 Mb or 79% of the Nipponbare genome. Sequence alignments with a PHRED quality value of 30 covered 12,319,100 bp (3%) of the total rice genome. A total of 80,127 sites differed in the corresponding regions in Nipponbare and Kasalath. The frequency of SNPs varied between chromosomes (0.53–0.78%). Insertions and deletions were also detected. The ratio of small insertion/deletion site nucleotides (1–14 bases) against the alignment length (0.20–0.27%) was similar among the different chromosomes, and there was no preference for the direction of insertions or deletions. The main patterns of base substitutions observed between Nipponbare and Kasalath are shown in Supplementar Table 16. Transitions (70%) were the most prominent substitutions; this is a substantially higher fraction than found between *Arabidopsis* ecotypes Columbia and Landsberg *erecta*³².

Class 1 simple sequence repeats in the rice genome

Class 1 simple sequence repeats (SSRs) are perfect repeats >20 nucleotides in length⁴⁵ that behave as hypervariable loci, providing a rich source of markers for use in genetics and breeding. A total of 18,828 Class 1 di-, tri- and tetra-nucleotide SSRs, representing 47 distinctive motif families, were identified and annotated on the rice genome (Supplementar Fig. 9). Supplementar Table 17 provides information about the physical positions of all Class 1 SSRs in relation to widely used restriction-fragment length polymorphisms (RFLPs)^{16,46} and previously published SSRs⁴⁵. There was an average of 51 hypervariable SSRs per Mb, with the highest density of markers occurring on chromosome 3 (55.8 SSR Mb⁻¹) and the lowest occurring on chromosome 4 (41.0 SSR Mb⁻¹). A summary of information about the Class 1 SSRs identified in the rice pseudomolecules appears

Table 2 | Size of each chromosome based on sequence data and estimated gaps

Chr	Sequenced bases (bp)	Gaps on arm regions No.	Length (Mb)	Telomeric gaps* (Mb)	Centromeric gap† (Mb)	rDNA‡ (Mb)	Total (Mb)	Coverage§ (%)	Coverage (%)
1	43,260,640	5	0.33	0.06	1.40		45.05	99.1	96.0
2	35,954,074	3	0.10	0.01	0.72		36.78	99.7	97.7
3	36,189,985	4	0.96	0.04	0.18		37.37	97.3	96.8
4	35,489,479	3	0.46	0.20			36.15	98.7	98.2
5	29,733,216	6	0.22	0.05			30.00	99.3	99.1
6	30,731,386	1	0.02	0.03	0.82		31.60	99.8	97.2
7	29,643,843	1	0.31	0.01	0.32		30.28	98.9	97.9
8	28,434,680	1	0.09	0.05			28.57	99.7	99.5
9	22,692,709	4	0.13	0.14	0.62	6.95	30.53	98.8	74.3
10	22,683,701	4	0.68	0.13	0.47		23.96	96.6	94.7
11	28,357,783	4	0.21	0.04	1.90	0.25	30.76	99.1	92.2
12	27,561,960	0	0.00	0.05	0.16		27.77	99.8	99.2
All	370,733,456	36	3.51	0.81	6.59	7.20	388.82	98.9	95.3

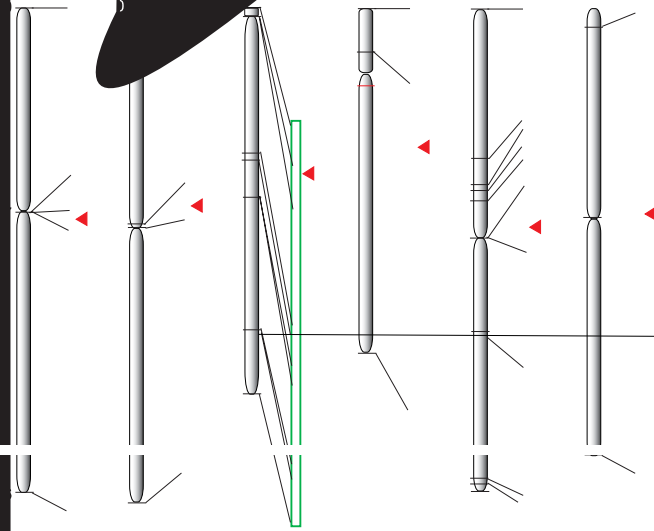
* Estimated length including the telomeres, calculated with the average value of 3.2 kb for each chromosome²⁴.

† Estimated length of centromere-specific CentO repeats on each chromosome²⁶.

‡ Represents the estimated length of the 17S–5.8S–25S rDNA cluster on Chr 9 (ref. 35) and the 5S cluster on Chr 11 (ref. 24).

§ Coverage of the pseudomolecules for the euchromatic regions in each chromosome.

|| Coverage of the pseudomolecules over the full length of each chromosome.



and elsewhere will accelerate marker-assisted breeding and positional cloning, facilitating advances in rice improvement.

The syntenic relationships between rice and the cereal grasses have long been recognized⁴. Comparing genome organization, genes and intergenic regions between cereal species will permit identification of regions that are highly conserved or rapidly evolving. Such regions are expected to yield crucial insights into genome evolution, speciation and domestication.

METHODS

Physical map and sequencing. Nine genomic libraries from *Oryza sativa* ssp. *japonica* cultivar Nipponbare were used to establish the physical map of rice chromosomes by polymerase chain reaction (PCR) screening¹⁹, fingerprinting²⁰ and end-sequencing²¹. The PAC, BAC and fosmid clones on the physical map were subjected to random shearing and shotgun sequencing to tenfold redundancy, using both universal primers and the dE-terminator or dE-primer methods. The sequences were assembled using PHRED (http://www.genome.washington.edu/UWGC/anal_sistools/Phred.cfm) and PHRAP (http://www.genome.washington.edu/UWGC/anal_sistools/Phrap.cfm) software packages or using the TIGR Assembler (http://www.tigr.org/software/assembler/).

Sequence gaps were resolved by full sequencing of gap-bridge clones, PCR fragments or direct sequencing of BACs. Sequence ambiguities (indicated by PHRAP scores less than 30) were resolved by confirming the sequence data using alternative chemistries or different polymerases. We empirically determined that a PHRAP score of 30 or above exceeds the standard of less than one error in 10,000 bp. BAC and PAC assemblies were tested for accuracy by comparing computationally derived fingerprint patterns with experimentally determined patterns of restriction enzyme digests. Sequence quality was also evaluated by comparing independently obtained overlapping sequences.

Small physical gaps were filled by long-range PCR. Remaining physical gaps were measured using fluorescence *in situ* hybridization analysis. We used the length of CentO arrays²⁶ to estimate the size of each of the remaining centromere gaps.

Annotation and bioinformatics. Gene models were predicted using FGENESH (http://www.softberry.com/berry.phtml?topic=fgenesh) using the monocot trained matrix on the native and repeat-masked pseudomolecules. Gene models with incomplete open reading frames, those encoding proteins of less than 50 amino acids, or those corresponding to organellar DNA were omitted from the final set. The coordinates of transposable elements, excluding MITEs (miniature inverted-repeat transposable elements), were used to mask the pseudomolecules.

Conserved domain/motif searches and association with gene ontologies were performed using InterProScan (http://www.ebi.ac.uk/InterProScan/) in combination with the InterPro2Go program. For biological processes, the number of detected domains was re-calculated as number of non-redundant proteins.

The predicted rice proteome was searched using BLASTP against the proteomes of several model species for which a complete genome sequence and deduced protein set was available. Each rice chromosome was searched against the TIGR rice gene index (http://www.tigr.org/tdb/tgi/ogi/) and against gene index entries that aligned to gene models corresponding to expressed genes. In addition, vertebrate gene indices (http://www.tigr.org/tdb/tgi/) were searched

against the rice chromosomes, and gene index matches were recorded. We searched the *Oryza sativa* ssp. *japonica* cv. Nipponbare collection of full-length cDNAs (ftp://cdna01.dna.affrc.go.jp/pub/data/), after first removing the transposable-element-related sequences, against the FGENESH models.

Gene models with rice full-length cDNA, EST or cereal EST matches but without identifiable homologues in the *Arabidopsis* genome were searched for conserved domains/motifs using InterProScan, and for homologues in the Swiss-Prot database (http://us.expasy.org/sprot/) using BLASTP. All proteins with positive blast matches were further compared with the nr database (http://www.ncbi.nlm.nih.gov/blast/html/blastcgihelp.html#protein_databases), using BLASTP to eliminate truncated proteins and those with matches to other dicots.

Tandem gene families. The rice genome was subjected to a BLASTP search as previously described³². The search was also performed by permitting more than one unrelated gene within the arrays, and the limit of the search was set to 5-Mb intervals to exclude large chromosomal duplications.

Non-coding RNAs. Transfer-RNA genes were detected by the program tRNA-scan SE (http://www.genetics.wustl.edu/edd/tRNAscan-SE/). The miRNA register in the Rfam database (http://www.sanger.ac.uk/Software/Rfam/) was used as a reference database for miRNAs. In addition, experimentally validated miRNAs of other species, excluding *Arabidopsis* miRNAs, were used for BLASTN queries against the pseudomolecules. Spliceosomal and snoRNAs were retrieved from the Rfam database and used for queries. BLASTN was used to find the location of snoRNAs and spliceosomal RNAs in the pseudomolecules.

Organellar insertions. *Oryza sativa* ssp. *japonica* Nipponbare chloroplast (GenBank NC_001320) and mitochondrial (GenBank BA000029) sequences were aligned with the pseudomolecules using BLASTN and MUMmer⁴⁹.

Transposable elements. The TIGR *Oryza* Repeat Database, together with other published and unpublished rice transposable element sequences, was used to create RTEdb (a rice transposable element database)⁵⁰ and determine transposable element coordinates on the rice pseudomolecules. In the case of *hAT*, *IS256/Mutator*, *IS5/Tourist* and *IS630/Tc1/mariner* elements, family-specific profile hidden Markov models were applied using HMMER⁴¹ (http://hmm.wustl.edu/). The remaining superfamilies were annotated using RepeatMasker (http://www.repeatmasker.org/).

Tos17 insertions. Flanking sequences of transposed copies of 6,278 *Tos17* insertion lines were isolated by modified thermal asymmetric interlaced (TAIL)-PCR and suppression PCR, and screened against the pseudomolecule sequences.

SNP discovery. BAC clones from an *O. sativa* ssp. *indica* var. Kasalath BAC library were end-sequenced. Sequence reads were omitted if they contained more than 50% nucleotides of low quality or high similarity to known repeats. The remaining sequences were subjected to BLASTN analysis against the pseudomolecules. Gaps within the alignments were classified as small insertions/deletions.

SSR loci. The Simple Sequence Repeat Identification Tool (http://www.gramene.org/) was used to identify simple sequence repeat motifs, and the physical position of all Class I SSRs was recorded. The copy number of SSR markers was estimated using electronic (e)-PCR to determine the number of independent hits of primer pairs on the pseudomolecules.

Whole-genome shotgun assembly analysis. Contigs from the BGI 6.28 × whole genome assembly of *O. sativa* ssp. *indica* 93-11 (GenBank/DDBJ/EMBL accession number AAAA02000001 AAAA02050231) and the Sngenta 6 × whole genome assembly of *O. sativa* ssp. *japonica* cv. Nipponbare (AACV01000001 AACV01035047; ref. 48) were aligned with the pseudomolecules using MUMmer⁴⁹. The number of IRGSP Nipponbare full-length cDNA-supported gene models completely covered by the aligned contigs was tabulated. The 155-bp CentO consensus sequence was used for BLAST analysis against the 93-11 and Nipponbare whole-genome shotgun contigs, and the coordinates of the positive hits recorded. Locations of centromeres for each *indica* chromosome were obtained with the CentO sequence positions on the IRGSP pseudomolecule of the corresponding chromosome. A detailed comparison of the BGI-assembled and -mapped Sngenta contigs (AACV01000001 AACV01000070) and the 93-11 contigs (AAAA02000001 AAAA02000093) was obtained by BLAST analysis against the IRGSP chromosome 1 pseudomolecule.

Detailed procedures for the analyses described above can be found in the Supplemental Information.

Received 29 December 2004; accepted 25 May 2005.

- Peng, S., Cassman, K. G., Virmani, S. S., Sheehy, J. & Khush, G. S. Yield potential trends of tropical rice since the release of IR8 and the challenge of increasing rice yield potential. *Crop Sci.* **39**, 1552–1559 (1999).
- Peng, S. *et al.* Rice yields decline with higher night temperature from global warming. *Proc. Natl Acad. Sci. USA* **101**, 9971–9975 (2004).

Table 3 | Transposons in the rice genome

	Copy no. ($\times 10^3$)	Coverage (kb)	Fraction of genome (%)
Class I			
LINEs	9.6	4161.3	1.12
SINEs	1.8	209.9	0.06
Ty1/copia	11.6	14266.7	3.85
Ty3/gypsy	23.5	40363.3	10.90
Other class I	15.4	12733.3	3.43
Total class I	61.9	71734.4	19.35
Class II			
<i>hAT</i>	1.1	1405.9	0.38
CACTA	10.8	9987.3	2.69
<i>IS630/Tc1/mariner</i>	67.0	8388.3	2.26
<i>IS256/Mutator</i>	8.8	13485.7	3.64
<i>IS5/Tourist</i>	57.9	12095.8	3.26
Other class II	18.2	2703.6	0.73
Total class II	163.8	48066.6	12.96
Other TEs	23.6	6797.7	1.80
Total TEs	249.3	129019.3*	34.79

TE, transposable element.

* Total length; corrected for 2420.7 kb in overlaps of multiple, non-nested elements.

3. Sasaki, T. & Burr, B. International Rice Genome Sequencing Project: the effort to completely sequence the rice genome. *Curr. Opin. Plant Biol.* **3**, 138–141 (2000).
4. Moore, G., Devos, K. M., Wang, Z. & Gale, M. D. Cereal genome evolution: Grasses, line up and form a circle. *Curr. Biol.* **5**, 737–739 (1995).
5. Sasaki, T. *et al.* The genome sequence and structure of rice chromosome 1. *Nature* **420**, 312–316 (2002).
6. Feng, Q. *et al.* Sequence and analysis of rice chromosome 4. *Nature* **420**, 316–320 (2002).
7. Rice Chromosome 10 Sequencing Consortium, In-depth view of structure, activity, and evolution of rice chromosome 10. *Science* **300**, 1566–1569 (2003).
8. Wu, J. *et al.* Composition and structure of the centromeric region of rice chromosome 8. *Plant Cell* **16**, 967–976 (2004).
9. Zhang, Y. *et al.* Structural features of the rice chromosome 4 centromere. *Nucleic Acids Res.* **32**, 2023–2030 (2004).
10. Nagaki, K. *et al.* Sequencing of a rice centromere uncovers active genes. *Nature Genet.* **36**, 138–145 (2004).
11. Guyot, R. & Keller, B. Ancestral genome duplication in rice. *Genome* **47**, 610–614 (2004).
12. Simillion, C., Vandepoele, K., Saey, Y. & Van de Peer, Y. Building genomic profiles for uncovering segmental homology in the twilight zone. *Genome Res.* **14**, 1095–1106 (2004).
13. Paterson, A. H., Bowers, J. E. & Chapman, B. A. Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proc. Natl Acad. Sci. USA* **101**, 9903–9908 (2004).
14. Salse, J., Piegue, B., Cooke, R. & Delseny, M. New *in silico* insight into the synteny between rice (*Oryza sativa* L.) and maize (*Zea mays* L.) highlights reshuffling and identifies new duplications in the rice genome. *Plant J.* **38**, 396–409 (2004).
15. Lai, J. *et al.* Gene loss and movement in the maize genome. *Genome Res.*

International Rice Genome Sequencing Project (Participants are arranged by area of contribution and then by institution.)

Physical Maps and Sequencing: Rice Genome Research Program (RGP) Takashi Matsumoto¹, Jianzhong Wu¹, Hiroyuki Kanamori¹, Yuichi Katayose¹, Masaki Fujisawa¹, Nobukazu Namiki¹, Hiroshi Mizuno¹, Kimiko Yamamoto¹, Baltazar A. Antonio¹, Tomoya Baba¹, Katsumi Sakata¹, Yoshiaki Nagamura¹, Hiroyoshi Aoki¹, Koji Arikawa¹, Kohei Arita¹, Takahito Bito¹, Yoshino Chiden¹, Nahoko Fujitsuka¹, Rie Fukunaka¹, Masao Hamada¹, Chizuko Harada¹, Akiko Hayashi¹, Saori Hijishita¹, Mikiko Honda¹, Satomi Hosokawa¹, Yoko Ichikawa¹, Atsuko Itonuma¹, Masumi Iijima¹, Michiko Ikeda¹, Maiko Ikeno¹, Kazue Ito¹, Sachie Ito¹, Tomoko Ito¹, Yuichi Ito¹, Yukiyo Ito¹, Aki Iwabuchi¹, Kozue Kamiya¹, Wataru Karasawa¹, Kanako Kurita¹, Satoshi Katagiri¹, Ari Kikuta¹, Harumi Kobayashi¹, Noriko Kobayashi¹, Kayo Machita¹, Tomoko Maehara¹, Masatoshi Masukawa¹, Tatsumi Mizubayashi¹, Yoshiyuki Mukai¹, Hideki Nagasaki¹, Yuko Nagata¹, Shinji Naito¹, Marina Nakashima¹, Yuko Nakama¹, Yumi Nakamichi¹, Mari Nakamura¹, Ayano Meguro¹, Manami Negishi¹, Isamu Ohta¹, Tomoya Ohta¹, Masako Okamoto¹, Nozomi Ono¹, Shoko Saji¹, Miyuki Sakaguchi¹, Kumiko Sakai¹, Michie Shibata¹, Takanori Shimokawa¹, Jianyu Song¹, Yuka Takazaki¹, Kimihiro Terasawa¹, Mika Tsugane¹, Kumiko Tsuji¹, Shigenori Ueda¹, Kazunori Waki¹, Harumi Yamagata¹, Mayu Yamamoto¹, Shinichi Yamamoto¹, Hiroko Yamane¹, Shoji Yoshiki¹, Rie Yoshihara¹, Kazuko Yukawa¹, Huisun Zhong¹, Masahiro Yano¹, Takuji Sasaki (Principal Investigator)¹;

The Institute for Genomic Research (TIGR) Qiaoping Yuan², Shu Ouyang², Jia Liu², Kristine M. Jones², Kristen Gansberger², Kelly Moffat², Jessica Hill², Jayati Bera², Douglas Fadrosh², Shaohua Jin², Shivani Johri², Mary Kim², Larry Overton², Matthew Reardon², Tamara Tsitir², Hue Vuong², Bruce Weaver², Anne Cieccko², Luke Tallon², Jacqueline Jackson², Grace Pai², Susan Van Aken², Terry Utterback², Steve Reidmuller², Tamara Feldblyum², Joseph Hsiao², Victoria Zismann², Stacey Iobst², Aymeric R. de Vazeille², C. Robin Buell (Principal Investigator)²;

National Center for Gene Research Chinese Academy of Sciences (NCGR) Kai Ying³, Ying Li³, Tingting Lu³, Yuchen Huang³, Qiang Zhao³, Qi Feng³, Lei Zhang³, Jingjie Zhu³, Qijun Weng³, Jie Mu³, Yiqi Lu³, Danlin Fan³, Yilei Liu³, Jianping Guan³, Yujun Zhang³, Shuliang Yu³, Xiaohui Liu³, Yu Zhang³, Guofan Hong³, Bin Han (Principal Investigator)³;

Genoscope Nathalie Choinsne⁴, Nadia Demange⁴, Gisela Orjeda⁴, Sylvie Samain⁴, Laurence Cattolico⁴, Eric Pelletier⁴, Arnaud Couloux⁴, Beatrice Segurens⁴, Patrick Wincker⁴, Angelique D'Hont⁴, Claude Scarpelli⁴, Jean Weissenbach⁴, Marcel Salanoubat⁴, Francis Quetier (Principal Investigator)⁴;

Arizona Genomics Institute (AGI) and Arizona Genomics Computational Laboratory (AGCol) Yeisoo Yu⁶, Hye Ran Kim⁶, Teri Rambo⁶, Jennifer Currie⁶, Kristi Collura⁶, Meizhong Luo⁶, Tae-Jin Yang⁶, Jetty S. S. Ammiraju⁶, Friedrich Engler⁶, Carol Soderlund⁶, Rod A. Wing (Principal Investigator)⁶;

Cold Spring Harbor Laboratory (CSHL) Lance E. Palmer⁷, Melissa de la Bastide⁷, Lori Spiegel⁷, Lidia Nascimento⁷, Theresa Zutavern⁷, Andrew O'Shaughnessy⁷, Sujit Dike⁷, Neilay Dedhia⁷, Raymond Preston⁷, Vivekanand Balija⁷, W. Richard McCombie (Principal Investigator)⁷;

Academia Sinica Plant Genome Center (ASPGC) Teh-Yuan Chow⁸, Hong-Hwa Chen⁹, Mei-Chu Chung⁸, Ching-San Chen⁸, Jei-Fu Shaw⁸, Hong-Pang Wu⁸, Kwang-Jen Hsiao¹⁰, Ya-Ting Chao⁸, Mu-kuei Chu⁸, Chia-Hsiung Cheng⁸, Ai-Ling Hour⁸, Pei-Fang Lee⁸, Shu-Jen Lin⁸, Yao-Cheng Lin⁸, John-Yu Liou⁸, Shu-Mei Liu⁸, Yue-le Hsing (Principal Investigator)⁸;

Indian Initiative for Rice Genome Sequencing (IIRGS), University of Delhi South Campus (UDSC) S. Raghuvanshi¹¹, A. Mohanty¹¹, A. K. Bharti^{11,13}, A. Gaur¹¹, V. Gupta¹¹, D. Kumar¹¹, V. Ravi¹¹, S. Vijai¹¹, A. Kapur¹¹, Parul Khurana¹¹, Paramjit Khurana¹¹, J. P. Khurana¹¹, A. K. Tyagi (Principal Investigator)¹¹;

Indian Initiative for Rice Genome Sequencing (IIRGS), Indian Agricultural Research Institute (IARI) K. Gaikwad¹², A. Singh¹², V. Dalal¹², S. Srivastava¹², A. Dixit¹², A. K. Pal¹², I. A. Ghazi¹², M. Yadav¹², A. Pandit¹², A. Bhargava¹², K. Sureshbabu¹², K. Batra¹², T. R. Sharma¹², T. Mohapatra¹², N. K. Singh (Principal Investigator)¹²;

Plant Genome Initiative at Rutgers (PGIR) Joachim Messing (Principal Investigator)¹³, Amy Bronzino Nelson¹³, Galina Fuks¹³, Steve Kavchok¹³, Gladys Keizer¹³, Eric Linton Victor Llaca¹³, Rentao Song¹³, Bahattin Tanyolac¹³, Steve Young¹³;

Korea Rice Genome Research Program (KRGRP) Kim Ho-Il¹⁴, Jang Ho Hahn (Principal Investigator)¹⁴;

National Center for Genetic Engineering and Biotechnology (BIOTEC) G. Sangsakoo¹⁵, A. Vanavichit (Principal Investigator)¹⁵;

Brazilian Rice Genome Initiative (BRIGI) Luiz Anderson Teixeira de Mattos¹⁶, Paulo Dejalma Zimmer¹⁶, Gaspar Malone¹⁶, Odir Dellagostin¹⁶, Antonio Costa de Oliveira (Principal Investigator)¹⁶;

John Innes Centre (JIC) Michael Bevan¹⁷, Ian Bancroft¹⁷;

Washington University School of Medicine Genome Sequencing Center Pat Minx¹⁸, Holly Cordum¹⁸, Richard Wilson¹⁸;

University of Wisconsin-Madison Zhukuan Cheng¹⁹, Weiwei Jin¹⁹, Jiming Jiang¹⁹, Sally Ann Leong²⁰

Annotation and Analysis: Hisakazu Iwama²¹, Takashi Gojobori^{21,22}, Takeshi Itoh^{22,23}, Yoshihito Niimura²⁴, Yasuyuki Fujii²⁵, Takuya Habara²⁵, Hiroaki Sakai^{23,25}, Yoshiharu Sato²², Greg Wilson²⁶, Kiran Kumar²⁷, Susan McCouch²⁶, Nikoleta Juretic²⁸, Douglas Hoen²⁸, Stephen Wright²⁹, Richard Bruskiewich³⁰, Thomas Bureau²⁸, Akio Miyao²³, Hirohiko Hirochika²³, Tomotaro Nishikawa²³, Koh-ichi Kadowaki²³ & Masahiro Sugiura³¹

Coordination: Benjamin Burr³²

Affiliations for participants: ¹National Institute of Agrobiological Sciences/Institute of the Society for Techno-innovation of Agriculture, Forestry and Fisheries, 2-1-2 Kannondai, Tsukuba, Ibaraki 305-8602, Japan. ²The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, Maryland 20850, USA. ³Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences (CAS), 500 Caobao Road, Shanghai 200233, China. ⁴Centre National de Séquençage, INRA-URGV, and CNRS UMR-8030, 2, rue Gaston Crémieux, CP 5706, 91057 EVRY Cedex, France. ⁵UMR PIA, Cirad-Amis, TA40-03 avenue Agropolis, 34398 Montpellier Cedex 05, France. ⁶Department of Plant Sciences, BIO5 Institute, The University of Arizona, Tucson, Arizona 85721, USA. ⁷Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11723, USA. ⁸Institute of Botany, Academia Sinica, 128, Sec. 2, Yen-Chiu-Yuan Rd, Nankang, Taipei 11529, Taiwan. ⁹National Cheng Kung University, No. 1, Ta-Hsueh Road, Tainan 701, Taiwan. ¹⁰National Yang-Ming University, 155, Sec. 2, Li-Nong St, Peitou, Taipei 112, Taiwan. ¹¹Department of Plant Molecular Biology, University of Delhi South Campus, New Delhi 110021, India. ¹²National Research Centre on Plant Biotechnology, Indian Agricultural Research Institute, New Delhi 110012, India. ¹³Waksman Institute, Rutgers University, Piscataway, New Jersey 08854, USA. ¹⁴National Institute of Agricultural Science and Technology, RDA, Suwon, 441-707 Republic of Korea. ¹⁵Rice Gene Discovery Unit, Kasetsart University, Nakron Pathom 73140, Thailand. ¹⁶Centro de Genômica e Fitomelhoramento, UFPel, Pelotas, RS, I 96001-970, Brazil. ¹⁷John Innes Centre, Norwich Research Park, Colney, Norwich NR4 7UH, UK. ¹⁸Washington University Genome Sequencing Center, 3333 Forest Park Boulevard, St. Louis, Missouri 63108, USA. ¹⁹University of Wisconsin, Department of Horticulture, Madison, Wisconsin 53706, USA. ²⁰University of Wisconsin, Department of Plant Pathology, Madison, Wisconsin 53706, USA. ²¹Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, Mishima 411-8540, Japan. ²²Biological Information Research Center, National Institute of Advanced Industrial Science and Technology, Koto-ku, Tokyo 135-0064, Japan. ²³National Institute of Agrobiological Sciences, Tsukuba, Ibaraki 305-8602, Japan. ²⁴Medical Research Institute, Tokyo Medical and Dental University, Bunkyo-ku, Tokyo 113-8510, Japan. ²⁵Japan Biological Information Research Center, Japan Biological Informatics Consortium, Koto-ku, Tokyo 135-0064, Japan. ²⁶Plant Breeding Dept, Cornell University, Ithaca, New York 14850-1901, USA. ²⁷Cold Spring Harbor Laboratory, PO Box 100, 1 Bungtown Road, Cold Spring Harbor, New York 11724, USA. ²⁸Department of Biology, McGill University, 1205 Dr Penfield Avenue, Montreal, Quebec H3A 1B1, Canada. ²⁹Department of Biology, York University, 4700 Keele Street, Toronto, Ontario M3J 1P3, Canada. ³⁰Biometrics and Bioinformatics Unit, International Rice Research Institute, DAPO Box 7777, Metro Manila, Philippines. ³¹Graduate School of Natural Sciences, Nagoya City University, Nagoya 467-8501, Japan. ³²Biology Department, Brookhaven National Laboratory, Upton, New York 11973, USA.