ARTICLES

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 development[1]. Rice grown on the most productive irrigated land has achieved nearly maximum production with current strains[1]. Environmental degradation, including pollution, increase in night time temperature due to global warming[2], 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 transformation[3]. The discovery of extensive genome colinearity among the Poaceae[4] 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 syntenic 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 P1 artificial chromosome clones (BACs and PACs, respectively). This strategy used a high-density genetic map[16], expressed-sequence tags (ESTs)[17], yeast artificial chromosome (YAC)- and BAC-based physical maps[18–20], BAC-end sequences[21] 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.

Ninety-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% (Supplementary 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 cytometry[24], and 403 Mb on the basis of lengths of anchored BAC contigs and estimates of gap sizes[20]. 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 markers[19]; 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 arrayed 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 (Supplementary 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 (Supplementary 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 rice[27], and provide a powerful resource for improving gene model structure derived from *ab initio* gene finders[28]. Of the 37,544 non-transposable-element-related FGENESH models, 17,016 could be supported by a total of 25,636 full-length cDNAs (Supplementary 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 (Supplementary 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 (Supplementary 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 (Supplementary Table 7). Fifty-one per cent of the predicted proteins could be associated with a biological process (Supplementary 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 (Supplementary 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 (Supplementary 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 *copia*-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 regenerated[29]. 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 produced[30]. A total of 11,487 target loci were mapped on the 12 pseudomolecules (Supplementary Fig. 5), with at least one insertion detected in 3,243 genes. The density of *Tos17* insertions is higher in euchromatic regions of the genome[30], 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 repeats[32]. 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 glycine-rich protein with 27 copies and one encoding a TRAF/BTB domain protein with 48 copies[33]. 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 (Supplementary Fig. 6). Sixty five per cent of the tandem arrays with over 27 members, and 33% of all the arrays with over 10 members, contain protein kinase domains (Supplementary Table 9).

Non-coding RNA genes

The nucleolar organizer, 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 centromere[36], 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]. Supplementary Fig. 7 shows the distribution of these tRNA genes in each chromosome. Chromosome 4 has a single tRNA cluster[6], and chromosome 10 has two large clusters derived from inserted chloroplast DNA[7]. 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 RNA[37]. miRNAs have been predicted from *Arabidopsis*[38] and rice[39], and we mapped 158 miRNAs onto the rice pseudomolecules (Supplementary 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 (Supplementary Table 11).

Organellar 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 internalized, 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. Homology searches detected 421–453 chloroplast insertions and 909–1,191 mitochondrial insertions, depending upon the stringency adopted (Supplementary Fig. 8 and Supplementary 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 (Supplementary Fig. 8 and Supplementary 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 Wisconcin) 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 (Supplementary 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 query sequences and the use of profile hidden Markov models allow the identification of more divergent elements[41] 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 IS*5*/*Tourist* and IS*630*/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 *h*AT, CACTA, IS*256*/*Mutator*, IS*5*/*Tourist*, and IS*630*/Tc1/*mariner* superfamilies, outnumber class I elements, which include long terminal-repeat (LTR) retrotransposons (Ty1/*copia*, Ty3/*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 IS*5*/*Tourist* and IS*630*/Tc1/*mariner* elements. The inverse is the case for maize, for which class I elements outnumber class II elements[42]. 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 (Supplementary Table 14). In contrast, members of some transposon superfamilies, including IS5/*Tourist*, IS*630*/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 population[16]. 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 (Supplementary 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 Supplementary 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 length[45] 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 (Supplementary Fig. 9). Supplementary 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 SSRs[45]. 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.	Gaps on arm regions 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 chromosome[24].
†Estimated length of centromere-specific CentO repeats on each chromosome[26].
‡Represents the estimated length of the17S–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.

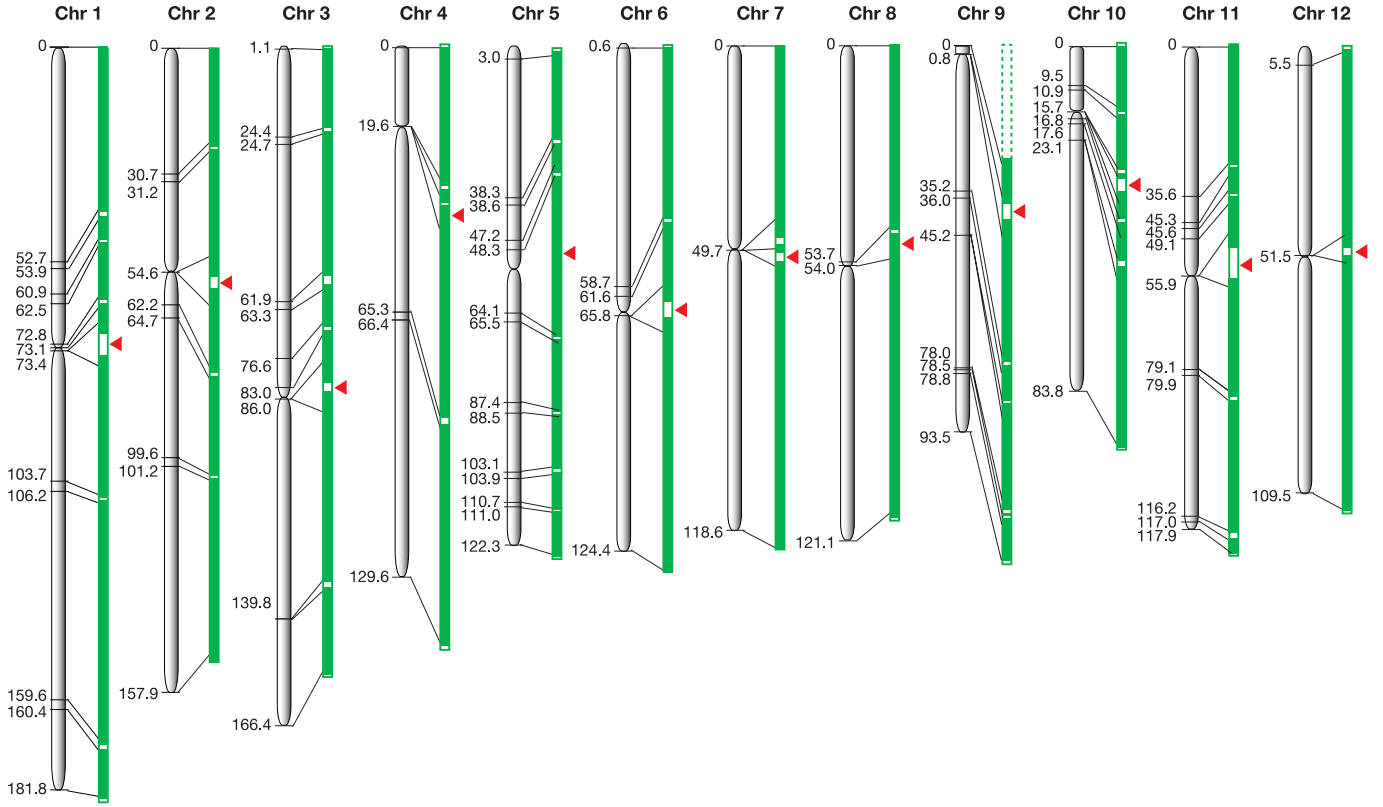


Figure 1 | Maps of the twelve rice chromosomes. For each chromosome (Chr 1–12), the genetic map is shown on the left and the PAC/BAC contigs on the right. The position of markers flanking the PAC/BAC contigs (green) is indicated on the genetic map. Physical gaps are shown in white and the nucleolar organizer on chromosome 9 is represented with a dotted green line. Constrictions in the genetic maps and arrowheads to the right of physical maps represent the chromosomal positions of centromeres for which rice CentO satellites are sequenced. The maps are scaled to genetic distances in centimorgans (cM) and the physical maps are depicted in relative physical lengths. Please refer to Table 2 for estimated lengths of the chromosomes.

in Supplementary Table 18. Several thousand of these SSRs have already been shown to amplify well and be polymorphic in a panel of diverse cultivars[45], and thus are of immediate use for genetic analysis.

Genome-wide comparison of draft versus finished sequences

Two whole-genome shotgun assemblies of draft-quality rice sequence have been published[23,47], and reassemblies of both have just appeared[48]. One of these is an assembly of 6.28 × coverage of *O. sativa* ssp. *indica* cv. 93-11. The second sequence is a ~6 × coverage of *O. sativa* ssp. *japonica* cv. Nipponbare[23,48]. These assemblies predict genome sizes of 433 Mb for *japonica* and 466 Mb for. *indica*, which differ from our estimation of a 389 Mb *japonica* genome. Contigs from the whole-genome shotgun assembly of 93-11 and Nipponbare[48] were aligned with the IRGSP pseudomolecules. Non-redundant coverage of the pseudomolecules by the *indica* assembly varied from 78% for chromosome 3 to 59% for chromosome 12, with an overall coverage of 69% (Supplementary Table 19). When genes supported by full-length cDNA coverage were aligned to the covered regions, we found that 68.3% were completely covered by the *indica* sequences. The average size of the *indica* contigs is 8.2 kb, so it is not surprising that many did not completely cover the gene models defined here. The coverage of the Nipponbare whole-genome shotgun assembly varied from 68–82%, with an overall coverage of 78% of the genome, and 75.3% of the full-length cDNAs supported gene models.

We undertook a detailed comparison of the first Mb of these assemblies on 1S (the short arm of chromosome 1) with the IRGSP chromosome 1 (Supplementary Fig. 10 and Supplementary Table 20). The numbers from this comparison agree with the whole-genome comparison described above. In addition, we observed that a substantial portion of the contigs from each assembly were non-homologous, misaligned or provided duplicate coverage. Indeed, the whole-genome shotgun assembly differed by 0.05% base-pair mismatches for the two aligned regions from the same Nipponbare cultivar. The two assemblies were further examined for the presence of the CentO sequence (Supplementary Table 21). Sixty-eight per cent of the copies observed in the 93-11 assembly and 32% of the CentO-containing contigs in the whole-genome shotgun Nipponbare assembly were found outside the centromeric regions. In contrast, the CentO repeats were restricted to the centromeric regions in the IRGSP pseudomolecules. It is unlikely that there are dispersed centromeres in *indica* rice; misassembly of the whole-genome shotgun sequences is a more likely explanation for dispersed CentO repeats. These observations indicate that the draft sequences, although providing a useful preliminary survey of the genome, might not be adequate for gene annotation, functional genomics or the identification of genes underlying agronomic traits.

Concluding remarks

The attainment of a complete and accurate map-based sequence for rice is compelling. We now have a blueprint for all of the rice chromosomes. We know, with a high level of confidence, the distribution and location of all the main components—the genes, repetitive sequences and centromeres. Substantial portions of the map-based sequence have been in public databases for some time, and the availability of provisional rice pseudomolecules based on this sequence has provided the scientific community with numerous opportunities to evaluate the genome, as indicated by the number of publications in rice biology and genetics over the past few years. Furthermore, the wealth of SNP and SSR information provided here

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 recognized[4]. 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) screening[19], fingerprinting[20] and end-sequencing[21]. 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 dye-terminator or dye-primer methods. The sequences were assembled using PHRED (http://www.genome.washington.edu/UWGC/analysistools/Phred.cfm) and PHRAP (http://www.genome.washington.edu/UWGC/analysistools/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 arrays[26] 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, five cereal 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 described[32]. 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 tRNAscan SE (http://www.genetics.wustl.edu/eddy/tRNAscan-SE/). The miRNA registry 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 MUMmer[49].

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 *h*AT, IS*256*/*Mutator*, IS*5*/*Tourist* and IS*630*/Tc1/*mariner* elements, family-specific profile hidden Markov models were applied using HMMER[41] (http://hmmer.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 1 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 Syngenta 6 × whole genome assembly of *O. sativa* ssp. *japonica* cv. Nipponbare (AACV01000001–AACV01035047; ref. 48) were aligned with the pseudomolecules using MUMmer[49]. 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 Syngenta 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 Supplementary Information.

Table 3 | Transposons in the rice genome

	Copy no. (× 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			
*h*AT	1.1	1405.9	0.38
CACTA	10.8	9987.3	2.69
IS*630*/Tc1/*mariner*	67.0	8388.3	2.26
IS*256*/*Mutator*	8.8	13485.7	3.64
IS*5*/*Tourist*	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.

Received 29 December 2004; accepted 25 May 2005.

1. 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).
2. Peng, S. *et al.* Rice yields decline with higher night temperature from global warming. *Proc. Natl Acad. Sci. USA* **101,** 9971–9975 (2004).

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., Saeys, 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., Piegu, 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.* **14,** 1924–1931 (2004).
16. Harushima, Y. *et al.* A high-density rice genetic linkage map with 2275 markers using a single F_2 population. *Genetics* **148,** 479–494 (1998).
17. Yamamoto, K. & Sasaki, T. Large-scale EST sequencing in rice. *Plant Mol. Biol.* **35,** 135–144 (1997).
18. Saji, S. *et al.* A physical map with yeast artificial chromosome (YAC) clones covering 63% of the 12 rice chromosomes. *Genome* **44,** 32–37 (2001).
19. Wu, J. *et al.* A comprehensive rice transcript map containing 6591 expressed sequence tag sites. *Plant Cell* **14,** 525–535 (2002).
20. Chen, M. *et al.* An integrated physical and genetic map of the rice genome. *Plant Cell* **14,** 537–545 (2002).
21. Mao, L. *et al.* Rice transposable elements: a survey of 73,000 sequence-tagged-connectors. *Genome Res.* **10,** 982–990 (2000).
22. Barry, G. F. The use of the Monsanto draft rice genome sequence in research. *Plant Physiol.* **125,** 1164–1165 (2001).
23. Goff, S. A. *et al.* A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science* **296,** 92–100 (2002).
24. Ohmido, N., Kijima, K., Akiyama, Y., de Jong, J. H. & Fukui, K. Quantification of total genomic DNA and selected repetitive sequences reveals concurrent changes in different DNA families in *indica* and *japonica* rice. *Mol. Gen. Genet.* **263,** 388–394 (2000).
25. Dong, F. *et al.* Rice (*Oryza sativa*) centromeric regions consist of complex DNA. *Proc. Natl Acad. Sci. USA* **95,** 8135–8140 (1998).
26. Cheng, Z. *et al.* Functional rice centromeres are marked by a satellite repeat and a centromere-specific retrotransposon. *Plant Cell* **14,** 1691–1704 (2002).
27. Kikuchi, S. *et al.* Collection, mapping, and annotation of over 28,000 cDNA clones from japonica rice. *Science* **301,** 376–379 (2003).
28. Castelli, V. *et al.* Whole genome sequence comparisons and "full-length" cDNA sequences: a combined approach to evaluate and improve *Arabidopsis* genome annotation. *Genome Res.* **14,** 406–413 (2004).
29. Hirochika, H., Sugimoto, K., Otsuki, Y., Tsugawa, H. & Kanda, M. Retrotransposons of rice involved in mutations induced by tissue culture. *Proc. Natl Acad. Sci. USA* **93,** 7783–7788 (1996).
30. Miyao, A. *et al.* Target site specificity of the *Tos17* retrotransposon shows a preference for insertion within genes and against insertion in retrotransposon-rich regions of the genome. *Plant Cell* **15,** 1771–1780 (2003).
31. Alonso, J. M. *et al.* Genome-wide insertional mutagenesis of *Arabidopsis thaliana*. *Science* **301,** 653–657 (2003).
32. Arabidopsis Genome Initiative, Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408,** 796–815 (2000).
33. Song, R., Llaca, V. & Messing, J. Mosaic organization of orthologous sequences in grass genomes. *Genome Res.* **12,** 1549–1555 (2002).
34. Shishido, R., Sano, Y. & Fukui, K. Ribosomal DNAs: an exception to the conservation of gene order in rice genomes. *Mol. Gen. Genet.* **263,** 586–591 (2000).
35. Oono, K. & Sugiura, M. Heterogeneity of the ribosomal RNA gene clusters in rice. *Chromosoma* **76,** 85–89 (1980).
36. Kamisugi, Y. *et al.* Physical mapping of the 5S ribosomal RNA genes on rice chromosome 11. *Mol. Gen. Genet.* **245,** 133–138 (1994).
37. Bartel, D. P. MicroRNAs: Genomics, biogenesis, mechanism, and function. *Cell* **116,** 281–297 (2004).
38. Wang, X. J., Reyes, J. L., Chua, N. H. & Gaasterland, T. Prediction and identification of *Arabidopsis thaliana* microRNAs and their mRNA targets. *Genome Biol.* **5,** R65 (2004).
39. Wang, J. F., Zhou, H., Chen, Y. Q., Luo, Q. J. & Qu, L. H. Identification of 20 microRNAs from *Oryza sativa*. *Nucleic Acids Res.* **32,** 1688–1695 (2004).
40. Turcotte, K., Srinivasan, S. & Bureau, T. Survey of transposable elements from rice genomic sequences. *Plant J.* **25,** 169–179 (2001).
41. Eddy, S. R. Profile hidden Markov models. *Bioinformatics* **14,** 755–763 (1998).
42. Messing, J. *et al.* Sequence composition and genome organization of maize. *Proc. Natl Acad. Sci. USA* **101,** 14349–14354 (2004).
43. Shen, Y. J. *et al.* Development of genome-wide DNA polymorphism database for map-based cloning of rice genes. *Plant Physiol.* **135,** 1198–1205 (2004).
44. Feltus, F. A. *et al.* An SNP resource for rice genetics and breeding based on subspecies indica and japonica genome alignments. *Genome Res.* **14,** 1812–1819 (2004).
45. McCouch, S. R. *et al.* Development and mapping of 2240 new SSR markers for rice (*Oryza sativa* L.). *DNA Res.* **9,** 257–279 (2002).
46. Causse, M. A. *et al.* Saturated molecular map of the rice genome based on an interspecific backcross population. *Genetics* **138,** 1251–1274 (1994).
47. Yu, J. *et al.* A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science* **296,** 79–92 (2002).
48. Yu, J. *et al.* The genomes of *Oryza sativa*: A history of duplications. *PLoS Biol.* **3,** e38 (2005).
49. Delcher, A. L. *et al.* Alignment of whole genomes. *Nucleic Acids Res.* **27,** 2369–2376 (1999).
50. Juretic, N., Bureau, T. E. & Bruskiewich, R. M. Transposable element annotation of the rice genome. *Bioinformatics* **20,** 155–160 (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements Work at the RGP was supported by the Ministry of Agriculture, Forestry and Fisheries of Japan. Work at TIGR was supported by grants to C.R.B. from the USDA Cooperative State Research, Education and Extension Service–National Research Initiative, the National Science Foundation and the US Department of Energy. Work at the NCGR was supported by the Chinese Ministry of Science and Technology, the Chinese Academy of Sciences, the Shanghai Municipal Commission of Science and Technology, and the National Natural Science Foundation of China. Work at Genoscope was supported by le Ministère de la Recherche, France. Funding for the work at the AGI and AGCoL was provided by grants to R.A.W. and C.S. from the USDA Cooperative State Research, Education and Extension Service–National Research Initiative, the National Science Foundation, the US Department of Energy and the Rockefeller Foundation. Work at CSHL was supported by grants from the USDA Cooperative State Research, Education and Extension Service–National Research Initiative and from the National Science Foundation. Work at the ASPGC was supported by Academia Sinica, National Science Council, Council of Agriculture, and Institute of Botany, Academia Sinica. The IIRGS acknowledges the Department of Biotechnology, Government of India, for financial assistance and the Indian Council of Agricultural Research, New Delhi, for support. Work at Rice Gene Discovery was supported by BIOTECH and the Princess Sirindhorn's Plant Germplasm Conservation Initiative Program. Work at PGIR was supported by Rutgers University. The BRIGI was supported by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Financiadora de Estudos e Projetos - Ministério de Ciência e Tecnologia (FINEP-MCT), Fundação de Amparo a Pesquisa do Rio Grande do Sul (FAPERGS) and Universidade Federal de Pelotas (UFPel). Work at McGill and York Universities was supported by the National Science and Engineering Research Council of Canada and the Canadian International Development Agency. Funding for H.H. at the National Institute of Agrobiological Sciences was from the Ministry of Agriculture, Forestry, and Fisheries of Japan, and the Program for Promotion of Basic Research Activities for Innovative Biosciences. Funding at Brookhaven National Laboratory was from The Rockefeller Foundation and the Office of Basic Energy Science of the United States Department of Energy. We would like to thank G. Barry and S. Goff for their help in negotiating agreements that permitted the sharing of materials and sequence with the IRGSP. We also acknowledge the work of G. Barry, S. Goff and their colleagues in facilitating the transfer of sequence information and supporting data.

Author Information The genomic sequence is available under accession numbers AP008207–AP008218 in international databases (DDBJ, GenBank and EMBL). Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to Takuji Sasaki (tsasaki@nias.affrc.go.jp).

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 Matsumoto[1], Jianzhong Wu[1], Hiroyuki Kanamori[1], Yuichi Katayose[1], Masaki Fujisawa[1], Nobukazu Namiki[1], Hiroshi Mizuno[1], Kimiko Yamamoto[1], Baltazar A. Antonio[1], Tomoya Baba[1], Katsumi Sakata[1], Yoshiaki Nagamura[1], Hiroyoshi Aoki[1], Koji Arikawa[1], Kohei Arita[1], Takahito Bito[1], Yoshino Chiden[1], Nahoko Fujitsuka[1], Rie Fukunaka[1], Masao Hamada[1], Chizuko Harada[1], Akiko Hayashi[1], Saori Hijishita[1], Mikiko Honda[1], Satomi Hosokawa[1], Yoko Ichikawa[1], Atsuko Idonuma[1], Masumi Iijima[1], Michiko Ikeda[1], Maiko Ikeno[1], Kazue Ito[1], Sachie Ito[1], Tomoko Ito[1], Yuichi Ito[1], Yukiyo Ito[1], Aki Iwabuchi[1], Kozue Kamiya[1], Wataru Karasawa[1], Kanako Kurita[1], Satoshi Katagiri[1], Ari Kikuta[1], Harumi Kobayashi[1], Noriko Kobayashi[1], Kayo Machita[1], Tomoko Maehara[1], Masatoshi Masukawa[1], Tatsumi Mizubayashi[1], Yoshiyuki Mukai[1], Hideki Nagasaki[1], Yuko Nagata[1], Shinji Naito[1], Marina Nakashima[1], Yuko Nakama[1], Yumi Nakamichi[1], Mari Nakamura[1], Ayano Meguro[1], Manami Negishi[1], Isamu Ohta[1], Tomoya Ohta[1], Masako Okamoto[1], Nozomi Ono[1], Shoko Saji[1], Miyuki Sakaguchi[1], Kumiko Sakai[1], Michie Shibata[1], Takanori Shimokawa[1], Jianyu Song[1], Yuka Takazaki[1], Kimihiro Terasawa[1], Mika Tsugane[1], Kumiko Tsuji[1], Shigenori Ueda[1], Kazunori Waki[1], Harumi Yamagata[1], Mayu Yamamoto[1], Shinichi Yamamoto[1], Hiroko Yamane[1], Shoji Yoshiki[1], Rie Yoshihara[1], Kazuko Yukawa[1], Huisun Zhong[1], Masahiro Yano[1], Takuji Sasaki (Principal Investigator)[1]; **The Institute for Genomic Research (TIGR)** Qiaoping Yuan[2], Shu Ouyang[2], Jia Liu[2], Kristine M. Jones[2], Kristen Gansberger[2], Kelly Moffat[2], Jessica Hill[2], Jayati Bera[2], Douglas Fadrosh[2], Shaohua Jin[2], Shivani Johri[2], Mary Kim[2], Larry Overton[2], Matthew Reardon[2], Tamara Tsitrin[2], Hue Vuong[2], Bruce Weaver[2], Anne Ciecko[2], Luke Tallon[2], Jacqueline Jackson[2], Grace Pai[2], Susan Van Aken[2], Terry Utterback[2], Steve Reidmuller[2], Tamara Feldblyum[2], Joseph Hsiao[2], Victoria Zismann[2], Stacey Iobst[2], Aymeric R. de Vazeille[2], C. Robin Buell (Principal Investigator)[2]; **National Center for Gene Research Chinese Academy of Sciences (NCGR)** Kai Ying[3], Ying Li[3], Tingting Lu[3], Yuchen Huang[3], Qiang Zhao[3], Qi Feng[3], Lei Zhang[3], Jingjie Zhu[3], Qijun Weng[3], Jie Mu[3], Yiqi Lu[3], Danlin Fan[3], Yilei Liu[3], Jianping Guan[3], Yujun Zhang[3], Shuliang Yu[3], Xiaohui Liu[3], Yu Zhang[3], Guofan Hong[3], Bin Han (Principal Investigator)[3]; **Genoscope** Nathalie Choisne[4], Nadia Demange[4], Gisela Orjeda[4], Sylvie Samain[4], Laurence Cattolico[4], Eric Pelletier[4], Arnaud Couloux[4], Beatrice Segurens[4], Patrick Wincker[4], Angelique D'Hont[5], Claude Scarpelli[4], Jean Weissenbach[4], Marcel Salanoubat[4], Francis Quetier (Principal Investigator)[4]; **Arizona Genomics Institute (AGI) and Arizona Genomics Computational Laboratory (AGCoI)** Yeisoo Yu[6], Hye Ran Kim[6], Teri Rambo[6], Jennifer Currie[6], Kristi Collura[6], Meizhong Luo[6], Tae-Jin Yang[6], Jetty S. S. Ammiraju[6], Friedrich Engler[6], Carol Soderlund[6], Rod A. Wing (Principal Investigator)[6]; **Cold Spring Harbor Laboratory (CSHL)** Lance E. Palmer[7], Melissa de la Bastide[7], Lori Spiegel[7], Lidia Nascimento[7], Theresa Zutavern[7], Andrew O'Shaughnessy[7], Sujit Dike[7], Neilay Dedhia[7], Raymond Preston[7], Vivekanand Balija[7], W. Richard McCombie (Principal Investigator)[7]; **Academia Sinica Plant Genome Center (ASPGC)** Teh-Yuan Chow[8], Hong-Hwa Chen[9], Mei-Chu Chung[8], Ching-San Chen[8], Jei-Fu Shaw[8], Hong-Pang Wu[8], Kwang-Jen Hsiao[10], Ya-Ting Chao[8], Mu-kuei Chu[8], Chia-Hsiung Cheng[8], Ai-Ling Hour[8], Pei-Fang Lee[8], Shu-Jen Lin[8], Yao-Cheng Lin[8], John-Yu Liou[8], Shu-Mei Liu[8], Yue-Ie Hsing (Principal Investigator)[8]; **Indian Initiative for Rice Genome Sequencing (IIRGS), University of Delhi South Campus (UDSC)** S. Raghuvanshi[11], A. Mohanty[11], A. K. Bharti[11,13], A. Gaur[11], V. Gupta[11], D. Kumar[11], V. Ravi[11], S. Vij[11], A. Kapur[11], Parul Khurana[11], Paramjit Khurana[11], J. P. Khurana[11], A. K. Tyagi (Principal Investigator)[11]; **Indian Initiative for Rice Genome Sequencing (IIRGS), Indian Agricultural Research Institute (IARI)** K. Gaikwad[12], A. Singh[12], V. Dalal[12], S. Srivastava[12], A. Dixit[12], A. K. Pal[12], I. A. Ghazi[12], M. Yadav[12], A. Pandit[12], A. Bhargava[12], K. Sureshbabu[12], K. Batra[12], T. R. Sharma[12], T. Mohapatra[12], N. K. Singh (Principal Investigator)[12]; **Plant Genome Initiative at Rutgers (PGIR)** Joachim Messing (Principal Investigator)[13], Amy Bronzino Nelson[13], Galina Fuks[13], Steve Kavchok[13], Gladys Keizer[13], Eric Linton Victor Llaca[13], Rentao Song[13], Bahattin Tanyolac[13], Steve Young[13]; **Korea Rice Genome Research Program (KRGRP)** Kim Ho-Il[14], Jang Ho Hahn (Principal Investigator)[14]; **National Center for Genetic Engineering and Biotechnology (BIOTEC)** G. Sangsakoo[15], A. Vanavichit (Principal Investigator)[15]; **Brazilian Rice Genome Initiative (BRIGI)** Luiz Anderson Teixeira de Mattos[16], Paulo Dejalma Zimmer[16], Gaspar Malone[16], Odir Dellagostin[16], Antonio Costa de Oliveira (Principal Investigator)[16]; **John Innes Centre (JIC)** Michael Bevan[17], Ian Bancroft[17]; **Washington University School of Medicine Genome Sequencing Center** Pat Minx[18], Holly Cordum[18], Richard Wilson[18]; **University of Wisconsin–Madison** Zhukuan Cheng[19], Weiwei Jin[19], Jiming Jiang[19], Sally Ann Leong[20]

Annotation and Analysis: Hisakazu Iwama[21], Takashi Gojobori[21,22], Takeshi Itoh[22,23], Yoshihito Niimura[24], Yasuyuki Fujii[25], Takuya Habara[25], Hiroaki Sakai[23,25], Yoshiharu Sato[22], Greg Wilson[26], Kiran Kumar[27], Susan McCouch[26], Nikoleta Juretic[28], Douglas Hoen[28], Stephen Wright[29], Richard Bruskiewich[30], Thomas Bureau[28], Akio Miyao[23], Hirohiko Hirochika[23], Tomotaro Nishikawa[23], Koh-ichi Kadowaki[23] & Masahiro Sugiura[31]

Coordination: Benjamin Burr[32]

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

Supplementary Information: Methods and Additional Results

Sequencing and physical map construction

The IRGSP adopted the hierarchical shotgun method for sequencing the rice genome. This strategy utilized nine genomic libraries from the *Oryza sativa* ssp. *japonica* cultivar Nipponbare to establish the physical map of rice and to aid in gap-filling (Supplementary Table 1). These included a P1-derived artificial chromosome (PAC) and three bacterial artificial chromosome (BAC) BAC libraries[1,2].During the course of sequencing Monsanto[3] and Syngenta[4] donated their draft sequences of the Nipponbare genome to the IRGSP. Monsanto also contributed BACs and collaborated in the construction of the BAC-based physical map. Monsanto's 5X draft sequenced BACs (638) underlie approximately 18% of the current sequence. Syngenta contigs were used for extending contigs and filling both physical and sequence gaps. Two complementary strategies were used to establish an accurate physical map. The RGP constructed a transcript map[5] with 6,591 STS/EST markers derived mostly from 3'UTR sequences of rice cDNAs. These markers were used to associate PAC/BAC clones with specific regions of the rice genome. Two of the BAC libraries were fingerprinted[2] and BAC-end sequenced by CUGI/AGI/AGCoL[6]. Using restriction digests of BAC DNAs, fingerprint contigs were also constructed and assembled by FPC[7]. Contigs were anchored to the map by probing clones with fragments or overgos derived from 603 mapped markers, and STCs were screened for sequences that matched these markers. Both the transcript map (http://rgp.dna.affrc.go.jp/publicdata/estmap2001/index.html) and FPC contig information (http://www.genome.arizona.edu/fpc/rice/) are available. The combined BAC/PAC libraries consist of a total of 212,160 clones and provide 60-fold coverage of the rice genome.

A sequence-ready physical map was constructed by first screening PAC/BAC libraries with either genetic markers or mapped ESTs in order to select seed clones[8]. Draft sequences of the seed clones were used to search for minimally overlapping clones from the BAC-end sequence database. FPC contigs with end-sequences that matched the sequences of seed BACs were added to the physical maps. This procedure was repeated several times to facilitate extension of the contigs. We used the PCR screening method[8] to search for clones that filled the remaining gaps. Additionally, two 10 kb insert genomic libraries and a 40 kb fosmid library were also constructed and utilized as an additional resource for gap-filling clones[9]. Finally, long range PCR was used for filling physical gaps below 40 kb. The PAC/BAC clones on the physical map were subjected to shotgun sequencing[10] using both universal primers[11] and the dye-terminator or dye-primer methods. Typically, 10-fold shotgun sequence redundancy was produced by random shearing of each PAC/BAC clone (3,840 sequences from 1,920 subclones with a size range of 2-8 kbp). For Monsanto clones, we complemented the draft sequence (5-fold redundancy) with a 5-fold overlapping sequence to produce a 10-fold sequence. The sequences were assembled by PHRED (http://www.genome.washington.edu/UWGC/analysistools/Phred.cfm) and PHRAP (http://www.genome.washington.edu/UWGC/analysistools/Phrap.cfm) software packages or with 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 low PHRAP scores were resolved by confirming the sequence data using alternative chemistries or different polymerases. The finished assemblies were verified by comparing sizes of virtual restriction digests with the experimental data. Standards for sequence quality

and annotation are described in the IRGSP Guidelines (http://demeter.bio.bnl.gov/Guidelines.html).

A small percentage of clones were identified and/or finished after construction of the pseudomolecules described in this manuscript. Thus, in addition to the pseudomolecule sequences which have been deposited in Genbank/DDBJ/EMBL, we have additional clones and newly finished sequences that contributed to our analyses.

Size measurements for the remaining physical gaps were conducted in different sequencing centers of IRGSP using the method of fluorescence *in situ* hybridization (FISH) analysis. Gaps on chromosome arms were measured using Fiber-FISH technology. Fiber-FISH or pachytene FISH technologies were used to measure telomere and centromere gaps. BAC or PAC clones from the termini of each contig were used as DNA probes. Additionally, synthetic oligonucleotides (CCCTAAA)n or DNA from the clone pAtT4 of Arabidopsis was used as the telomere probe. The preparation of somatic metaphase chromosomes and meiotic pachytene chromosomes as well as the FISH procedures were as described previously[12]. DNA probes were labeled with biotin-dUTP, digoxigenin-dUTP, or fluorescein isothiocyanate-dUTP (Boehringer Mannheim). Chromosomes were counterstained with propidium iodide or 4',6-diamidino-2-phenylindole. Hybridization signals in two-color fiber-FISH were detected with a three-layer antibody detection system. All images were captured digitally under an Olympus BX51 or BX60 epifluorescence microscope (Tokyo, Japan) using a SenSys charge-coupled device camera (Roper Scientific, Tucson, AZ). Gap length was measured using the probed BAC or PAC clone as a reference size marker[12].

The pseudomolecules include the mapped centromeres on chromosomes 4, 5 and 8. The pseudomolecules contain a portion of the CentO region from both the short and long arms for the 9 remaining centromeres. This result then allows us to use the length of CentO arrays quantified previously[13] to estimate the size of each of the remaining centromere gaps.

Sequences for seven chromosome ends (Chr1S, 2S, 2L, 6L, 7S, 7L and 8S) include telomere-specific sequences (CCCTAAA repeats, 17-88 copies) and were obtained from mapped and sequenced fosmid clones. The sequence on Chr2S is included in the pseudomolecule and the remaining 6 complete sequences are under submission. Sizes for the remaining telomere gaps were measured with the fiber- or pachytene-FISH method using terminal BAC/PAC clones as well as the telomere repeats as DNA probes. A value of 3.2 kb is used as the estimated length of each telomere[14].

Quality control

Several phases of quality control were implemented throughout the project. First, BAC/PAC assemblies were tested for accuracy through comparison of computationally-derived fingerprint patterns with experimentally determined restriction enzyme digestion patterns. Second, sequence quality was evaluated by comparing overlapping sequences from BAC/PAC clones that were in advertently sequenced in independent laboratories. The BAC/PAC sequences were retrieved from NCBI and aligned by Sequencher (Gene Codes, Inc., Ann Arbor, MI, USA). The alignment was inspected and number of base discrepancies was counted. A total of 14 clone pairs containing 1,247,885 bp of overlapping sequence (Supplementary Table 2)

were examined. Ten base substitutions in the corresponding regions were detected which indicates a sequence accuracy of 99.9992%. We have also detected 134 insertion/deletion differences. Detailed inspection revealed that most of these discrepancies came from simple sequence repeats in which the unit number of repeats was undetermined or variable within or between the clones and therefore marked as "sequence ambiguous region". If we combine both types of discrepancy, the overall accuracy was calculated as 99.9885%. Third, the IRGSP conducted a quality control exercise patterned after similar exercises carried out during the human genome project. Briefly, random BACs sequenced by each group were selected from GenBank and the *in silico* restriction digest pattern for several enzymes was generated based on the submitted sequence. This was then compared to the pattern for those enzymes observed following *in vitro* digestion of a sample of the submitted BAC. Some anomalies were observed initially and many, upon detailed investigation, were found to be naming discrepancies between centers. These were resolved where appropriate. Overall, those centers whose submissions represent the vast majority of the submitted sequence had a very low (or in several cases, undetectable) error rate in this step in the ongoing quality control process. Some possible errors were detected and these clones are being further investigated and will be corrected in the public databases as appropriate.

Furthermore, additional computational analyses of the raw data and assemblies are underway to gain a better appreciation for any other types of errors that may exist so that these too may be corrected. It is expected that this process will be carried out by one of the groups (CSHL/AGI/WUGSC) that carried out most of the initial quality control exercise. As in any sequence such as this, the rice sequence is a work in progress and we will continue to refine it to the greatest extent possible. It should be noted that this type of detailed analysis cannot be carried out in a very efficient manner on whole genome assemblies.

Annotation and bioinformatic methods

Gene models were predicted using FGENESH (http://www.softberry.com/berry.phtml?topic=fgenesh) using the monocot trained matrix. Gene models with incomplete ORFs due to missing initiation or termination codons, gene models that encode for proteins of less than 50 amino acids, and gene models with ambiguous sequence (represented by Ns) were omitted from the final gene model set. Gene models that corresponded to organellar DNA or bacterial IS sequences with an e-value of < 1e-5 were removed. Models were predicted by FGENESH twice using the native and repeat-masked pseudomolecules. To mask the pseudomolecules for repetitive sequences, the coordinates of the transposable elements (see below) were used to mask the pseudomolecules. Because MITEs are frequently associated with genes, they were not masked. The cutoffs were IS630/Tc1/mariner less than 300 bp, IS5/Tourist less than 400 bp, and all other MITEs less than 400bp.

With the completion of the rice genome, we have improved our ability to identify TE-related genes. A total of 55,296 genes were identified on the unmasked pseudomolecules, resulting in an apparent density of one gene per 6.7 kbp, similar to densities reported for finished analyses of chromosomes 1, 4 and 10[15-17]. However, this number, as well as the densities reported in these chromosome manuscripts, report on total genes including those related to transposable elements (TE-related) which are captured as “genes” in the annotation process.

Conserved domain/motif searches and association with gene ontologies was performed using InterproScan (Gene Ontology Consortium, http://www.ebi.ac.uk/GO/index.html, InterproScan ver.3.3, ipscanDATA ver.8.0) in combination with the Interpro2Go program. Only those domains that appeared over 100 times are listed in Supplementary Fig. 3b. For biological processes the number of detected domains were recalculated as non-redundant proteins (Supplementary Fig. 3a).

The *Oryza stativa* ssp. *japonica* cv Nipponbare collection of full-length cDNAs (ftp://cdna01.dna.affrc.go.jp/pub/data/) contains 33,678 FLcDNA sequences that cluster into ~21,000 clusters. After removal of 911 transposable element-related sequences, the number of FL-cDNA sequences was 32,767. To determine coverage of the non-TE-related FGENESH models, we searched the FLcDNAs against the FGENESH models using BLAST. Using a minimum of 95% identity and variable length cutoffs, the number of models supported by a FL-cDNA were determined (SupplementaryTable 6). The lack of full concordance between the FGENESH models and all the FL-cDNAs can be attributed to several factors including presence of incorrect gene structure in the FGENESH model, alternative splice forms present in the FL-cDNA collection but absent in the FGENESH set, genes missed by FGENESH, presence of pseudogenes and non-coding RNAs in the FL-cDNA collection but absent in the FGENESH models, gaps in the pseudomolecules, sequence quality issues with the FL-cDNAs, and aberrant FL-cDNAs in the FL-cDNA collection.

To identify putative homologs in other species, the predicted rice proteome was searched using BLASTP (cutoff criterion of $e < 1.0^{-5}$) against the proteomes of several model species for which a complete genome sequence and deduced protein set was available (Supplementary Fig. 4). To identify expressed genes, each rice chromosome was searched against the TIGR rice gene index13 (Release 15; http://www.tigr.org/tdb/tgi/ogi/) and gene index entries that aligned (cutoff criteria of 95 % identity over 80 % of the length of each gene index entry) to a gene model were parsed out. The density of expressed genes in 100 kb windows was plotted in Supplementary Fig. 2. In addition to the rice gene index, five cereal gene indices (wheat Release 8.0: http://www.tigr.org/tdb/tgi/tagi/, maize Release 14.0: http://www.tigr.org/tdb/tgi/zmgi/, barley Release 8.0: http://www.tigr.org/tdb/tgi/hvgi/, sorghum Release 8.0: http://www.tigr.org/tdb/tgi/sbgi/, and Rye Release 3.0: http://www.tigr.org/tdb/tgi/ryegi/) were searched against the rice chromosomes and gene indices matches (cutoff criteria of 70% identity over 80% of the length of the gene index sequence) were parsed out (Supplementary Fig. 1). A separate analysis using 32,127 full-length cDNA sequences (ftp://cdna01.dna.affrc.go.jp/pub/data/) was performed using BLASTN with the threshold e value of 10^{-20}.

To identify possible cereal-specific proteins the translated sequences of the 2,859 gene models with rice full-length cDNA or EST or cereal EST matches but without identifiable homologs in the *Arabidopsis* genome were searched for conserved domain/motif using InterproScan and homologs in the Swiss-Prot database (www.expasy.org/sprot/) using BLASTP with a cutoff value of e<-20. All proteins with positive blast matches and all proteins with interesting domain matches were further compared with the nr database (http://www.ncbi.nlm.nih.gov/blast/html/blastcgihelp.html#protein_databases) using BLASTP to eliminate all truncated proteins and those with matches to other dicots. The

gene models were also searched directly with rice homologs of proteins presumed to be cereal-specific, in particular expansins and hydroxyproline-rich glycoproteins.

Tandem gene family methods

To facilitate a comparison of rice and *Arabidopsis*, the rice genome was subjected to a BLASTP search with the same threshold e-value of 10^{-20} as described[18]. However, 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

One 33 kb fosmid clone that contains four 17S-5.8S-25S unit repeats was completely sequenced. All of the units were orientated on the chromosome in the same direction.

Transfer-RNA genes were detected by the program tRNA-scan SE[19]. We used the miRNA registry[20] in the Rfam database (http://www.sanger.ac.uk/Software/Rfam/, Release 5.1,Dec. 2004) as the reference database. All 134 entries in the *Oryza sativa* miRNA database have high similarities with the rice genome (more than 96.5% coverage and 93.6%) by using BLASTN. This is reasonable because these are the predicted homologues of the Arabidopsis miRNA against the whole genome shotgun assembly, accession AAAA01000000. We re-mapped these sequences at 129 positions in the pseudomolecules (Supplementary Table 10).

Experimentally validated miRNAs of other species, excluding *Arabidopsis* miRNAs, were used for BLASTN queries against the pseudomolecules. BLAST hits of more than 18 bases with the mature (20-22nt) miRNA sequences were retrieved with 1 kb flanking sequence on each side. Using these 2 kb sequences, Ssearch[21] was then performed against the original mature miRNA sequences used for BLAST analysis. Only the paired hits (inverted repeats) with mature miRNAs were selected. Among them, only mature miRNA sequence pairs which meet the following conditions were recorded: 1) More than 80% of the mature miRNA sequence is in the alignment; 2). The sequence identity in the aligned sequence is more than 80%; and 3) The set of the inverted repeat sequences are not more than 280 bases apart (280 nt is the maximal length of all the miRNA precursor entries in this database). No distribution bias on the chromosomes and their positions were observed (data not shown). Thus, we could add 29 new putative miRNAs onto the pseudomolecules for a total of 158 miRNAs identified in the genome.

All the snRNA entries (U-RNAs and snoRNAs) were retrieved from Rfam database (version6.1, August 2004) and used for queries. BLASTN was used to find the location of snoRNAs and spliceosomal RNAs in the pseudomolecules. Only the hits with more than 80% sequence identity were mapped onto the pseudomolecules. There are 379 total hits from 65 kinds of snoRNAs, and 368 total hits for U1, U2, U4, U5, and U6 spliceosomal RNA sequences. Overlaps in chromosomal position among these hit results were removed. Numbers of the non-redundantly mapped snoRNAs and spliceosomal RNAs are shown in Supplementary Table 11.

Chloroplast and mitochondrial methods

Oryza sativa ssp. *japonica* Nipponbare chloroplast[22] and mitochondrial[23] sequences were aligned with the 12 Nipponbare nuclear chromosomes by using BLASTN[24]

version 2.2.6 with default conditions. A low stringency search by MUMmer[25] was achieved using version 3.0 (with Bn Bl 18 BL ; mgaps Bl 100 Bf .12 Bs600 ; combineMUMs Bx Be .50 Bw) on the direct strand and on the reverse strand. A high stringency search was then achieved by removing the SW extensions 5= to the first MUM and 3= to the last MUM for each combine MUMs result.

TE Methods and Supplementary Information

The TIGR *Oryza* Repeat Database (TIGR RDB; http://www.tigr.org/tigr-scripts/e2k1/rpStat.cgi?DB=Oryza) was used as the starting point to create RTEdb[26], a database of known rice transposons (http://www.biology.mcgill.ca/faculty/bureau/index.htm). Published rice transposon sequences that were absent from TIGR RDB were incorporated into RTEdb[27-40]. Several unpublished retrotransposons (C. Vitte, personal communication) and a large set of unpublished *Mutator*-like elements (Bureau, personal communication) were also added. The revised database contained 18,342 complete and partial elements.

RTEdb was used to determine the TE coordinates on the rice pseudomolecules. In the case of *h*AT, IS*256*/*Mutator*, IS*5*/*Tourist*, and IS*630*/Tc1/*mariner* superfamilies, family-specific profile hidden Markov models (HMMs) were built and queried against the rice genomic sequences using the HMMER software package[41] (http://hmmer.wustl.edu/). The remaining transposon superfamilies were annotated using RepeatMasker (http://www.repeatmasker.org). A set of Perl scripts was written to process and consolidate the results generated by HMMER and RepeatMasker.

TE coordinates were used to examine their relationship with gene models and other genomic features. Recombination rates were estimated using the genetic and physical positions of mapped genetic markers along each chromosome. Third order polynomials were fit to the relationship between physical and genetic distance for each chromosome. Recombination rates were estimated for each 100 kb region of each chromosome, by taking the derivative of each polynomial at the midpoint of the 100 kb bin. Centromeric regions were defined using the marker data, as the region of suppressed recombination surrounding the centromere.

Transposon distribution patterns can generally be explained by a strong effect of average element length: small elements show a significant positive correlation with recombination rates and gene density, and are under-represented in the centromere, while larger elements are over-represented in gene- and recombination-poor heterochromatin (Supplementary Table 14). This generates a positive correlation between average transposon length and relative abundance in centromeric regions ($r=0.543$, $p<0.05$). The effect of transposon length suggests that larger elements are more likely to have deleterious effects in euchromatic regions, either by interfering with gene function, or by the deleterious consequences of ectopic recombination[42], and are thus removed by purifying selection.

***Tos17* Methods**

Flanking sequences of transposed copies of *Tos17* in each line were isolated by a modified TAIL-PCR and suppression PCR method[43]. Pseudomolecule sequences were split into 20 kb fragments with 10 kb overlaps to make a BLAST database. A total of 48,309 flanking sequences derived from 6,278 lines were screened with BLASTN

against the overlap database using the Score Cluster System at the Computer Center of MAFFIN, Japan (http://www.affrc.go.jp/). The BLASTN analysis required identities >= 99% and permitted 3 bp mismatch or a frame-shift at the 5'-end of flanking sequences. A total of 11,487 target loci were mapped on twelve pseudomolecules. The number of target loci in each 100 kb interval was plotted as a histogram using a perl script with pdflib_pl module (http://www.pdflib.com/) (Supplementary Fig. 5).

SNP Discovery

Comparison of Nipponbare *vs*. Kasalath: The construction of an *O. sativa* ssp. *indica* var. Kasalath BAC library was previously described[1]. The library consisted of 47,194 clones with an average insert size of 133 kb. All BAC clones were end-sequenced and a total of 78,752 BAC-end sequences (STCs) were obtained[44]. A BLAST search against all the Kasalath STCs is available at http://rgp.dna.affrc.go.jp/blast/runblast.html. The Kasalath STCs were mapped *in silico* onto the 12 Nipponbare pseudomolecules to determine the global DNA-sequence polymorphism between these two cultivars. The vector sequences were masked prior to sequence comparison. If more than 50% of the total nucleotides were masked, then the sequence reads were omitted. Sequence entries with many ambiguous bases (PHRED quality values q<5 over 50% of the total nucleotides) were also removed. The qualifying sequences were searched against the TIGR repeat databases (http://www.tigr.org/tdb/e2k1/osa1/blastsearch.shtml). Sequences with high similarity to known repeats such as transposons, retrotransposons, centromere repeats, and other rice repeat sequences with e-values of less than 10^{-50} were removed. The remaining sequences were subjected to BLASTN (NCBI-BLAST) analysis with the low complexity option against the 12 pseudomolecules. Additional criteria to detect only the base changes in the precisely mapped area were set as follows: (i) A valid hit is identified with e-value less than 10^{-100} from BLAST analysis. (ii) More than 50% of the STCs are aligned and BLAST analysis shows over 90% identity. (iii) Both end-sequences (forward and reverse directions) of the clone are mapped in opposite directions and are no more than 200 kb apart. These criteria facilitated accurate *in silico* mapping of BAC clones. Only the paired hits that mapped to a single specific site within a chromosome were used for *in silico* mapping and calculations. The genome coverage of the mapped BAC Kasalath clones was calculated relative to the Nipponbare physical map as the sum of the distance between singlet BAC STCs plus the distance between external STCs of contigs expressed as a percentage of the total length of the pseudomolecules. SNPs were analyzed using only high quality polymorphic nucleotides (q>30). Ns or Xs in both aligned sequences were not included for calculation of SNPs. The gaps within the alignments were classified as small InDels.

A quality check was done for 32 randomly selected BAC ends that were resequenced. These STCs had previously indicated 101 SNPs when aligned with the Nipponbare sequence. All 101 SNP nucleotides were confirmed upon resequencing. The RGP previously showed empirically that a PHRED value greater than 30 was the equivalent to less than one error in 10,000 (http://demeter.bio.bnl.gov/clemson.html).

SSR Methods and Discussion

We used the Simple Sequence Repeat Identification Tool (SSRIT)[45] available at www.gramene.org to identify simple sequence repeat motifs. Supplementary Table 17 provides information about the physical position of all Class 1 SSRs in relationship to a set of widely used genetic markers previously published as RFLPs[46,47] or SSRs[48]. The

bp position at the center of each SSR motif or at the 3' end of a sequenced RFLP marker is used to tag the position of the marker along the pseudomolecule. This makes it possible to determine the relative order of mapped loci and facilitates calculations of cumulative physical map distance based on a running summary of distances between markers.

A total of 18,828 Class 1 di, tri and tetra-nucleotide SSRs, representing 47 distinctive motif families, were identified in the rice genome. Fifty-nine percent of all Class 1 SSRs belonged to one of the four di-nucleotide families, 27% were placed into the eleven tri-nucleotide families, and 14% were categorized into 32 tetra-nucleotide families. The three most abundant motif families were poly(AT)n (37% of the total), poly(GA)n (18%), and poly(CCG)n (10%) (Supplementary Fig. 9). The motif, poly(GC)n, was unusually rare, occurring only 19 times in the genome, compared to poly(AC)n, the next rarest di-nucleotide motif (782 times). The 19 instances of poly(GC) were distributed over 10 of the 12 chromosomes and showed no distinct clustering. While poly(CCG)n is highly abundant in the genome, there is apparent selection against poly(GC)n, presumably due to its ability to cause frameshift mutations in GC-rich genic regions. A motif family represents all related motifs, including frame shifts of the defined reference motif and their reverse complements, i.e., family (AGT) {AGT, GTA, TAG, ACT, CTA, TAC}.

Copy number of SSR markers was estimated using e-PCR to determine the number of independent hits of primer pairs on the pseudomolecules. Primer pairs were developed for 17,719 (94%) of the SSR loci and e-PCR was used to confirm the location and expected copy number of each SSR for which primers had been designed. Of these, a total of 16,959 (96%) showed a unique hit to the expected region of the genome and were determined to be single copy. In Supplementary Table 18, (*) following a marker reagent_ID, i.e., AUT26023(*), indicates that the primer pair shows multiple hits using e-PCR.

Information on SSRs by map location can be found at http://www.gramene.org/Oryza_sativa/. Over two thousand of these markers have been experimentally evaluated and found to reliably produce 90% amplification success under a single, standard set of PCR conditions[48].

Whole genome shotgun assembly analysis

The 50,231 contigs from the BGI 6.28X whole genome assembly of *O. sativa* ssp. *indica* 93-11 (AAAA02000001-AAAA02050231) and the 35,047 contigs of the Syngenta 6X whole genome assembly of *O. sativa* ssp. *japonica* cv. Nipponbare (AACV01000001-AACV01035047)[49] were aligned with the pseudomolecules using MUMmer[25]. The aligned Syngenta contigs were required to cover at least 95% of the collinear Nipponbare region with greater than 95% identity (Supplementary Table 19). A lower stringency of 50% coverage and 80% identity was required for the alignment BGI contigs (Supplementary Table 19). The number of IRGSP Nipponbare FL-cDNA supported gene models completely covered by the aligned contigs were tabulated. The 155-bp CentO consensus sequence was used for BLAST analysis against the 93-11 and Nipponbare WGS contigs and the coordinates of the positive hits recorded (Supplementary Table 21). General locations of centromeres for each *indica* chromosome were obtained by comparing the physical positions of the CentO-containing contigs mapped by BGI with the CentO sequence positions on the IRGSP pseudomolecule of the corresponding chromosome.

One might ask if sequence polymorphism between the *indica* and *japonica* rice makes this an unfair comparison. The level of SNP frequency between 93-11 and Nipponbare has been reported to be from less than 0.1%[50] to 0.37%[51] far below the 20% level of non-identity required to eliminate the matches. Han and Xue[52] compared 2.3 Mb of finished sequence from both Nipponbare and a different *indica* cultivar, Guangluai 4, on chromosome 4. In addition to finding SNPs they also documented insertions and deletions between the two sequences that amounted to almost 13% of the Nipponbare sequence.

A detailed comparison of the BGI assemblies with the IRGSP sequence was undertaken for the top portion of chromosome 1S. The BGI assembled and mapped Syngenta contigs (AACV01000001-AACV01000070) and the 93-11 contigs (AAAA02000001-AAAA02000093) were compared by BLAST analysis against the IRGSP chromosome 1 pseudomolecule (Supplementary Fig. 10; Supplementary Table 20). These contigs had been assigned in order on the short arm of chromosome 1 from the telomere region[49].

References

1. Baba, T. *et al.* Construction and characterization of rice genome libraries: PAC library of japonica variety, Nipponbare, and BAC library of indica variety, Kasalath. *Bull. Natl. Inst. Agrobiol. Resour. (Japan)* **14**, 41-51 (2000).
2. Chen, M. *et al.* An integrated physical and genetic map of the rice genome. *Plant Cell* **14**, 537-545 (2002).
3. Barry, G. F. The use of the Monsanto draft rice genome sequence in research. *Plant Physiol*. **125**, 1164-1165 (2001).
4. Goff, S. A. *et al.* A draft sequence of the rice genome (*Oryza sativa* L.ssp. *japonica*). *Science* **296**, 92-100 (2002).
5. Wu, J. *et al.* A comprehensive rice transcript map containing 6591 expressed sequence tag sites. *Plant Cell* **14**, 525-535 (2002).
6. Mao, L. *et al.* Rice transposable elements: a survey of 73,000 sequence-tagged-connectors (STCs). *Genome Res*. **10**, 982-990 (2000).
7. Soderlund, C., Longden, I. & Mott, R. FPC: a system for building contigs from restriction fingerprinted clones. *CABIOS* **13**, 523-535 (1997).
8. Wu, J. *et al.* Physical maps and recombination frequency of 6 rice chromosomes. *Plant J*. **36**, 720-730 (2003).
9. Yang, T. J. *et al.* Construction and utility of 10-kb libraries for efficient clone-gap closure for rice genome sequencing. *Theor. Appl. Genet*. **107**, 652-660 (2003).
10. Messing, J., Crea, R. & Seeburg, P. H. A system for shotgun DNA sequencing. *Nucleic Acids Res*. **9**, 309-321 (1981).
11. Vieira, J. & Messing, J. The pUC plasmids, an M13mp7 derived system for insertion mutagenesis and sequencing with synthetic universal primers. *Gene* **19**, 259-268 (1982).
12. Cheng, Z. K., Buell, C. R., Wing, R. A., Gu, M. H. & Jiang, J. Toward a cytological characterization of the rice genome. *Genome Res*. **11**, 2133-2141 (2001).
13. Cheng, Z. *et al.* Functional rice centromeres are marked by a satellite repeat and a centromere-specific retrotransposon. *Plant Cell* **14**, 1691-1704 (2002).

14. Ohmido, N, Kijima, K, Akiyama, Y., de Jong, J. H. & Fukui, K. Quantification of total genomic DNA and selected repetitive sequences reveals concurrent changes in different DNA families in indica and japonica rice. *Mol. Gen. Genet*. **263**, 388-394 (2000).
15. Sasaki, T. *et al*. The genome sequence and structure of rice chromosome 1. *Nature* **420**, 312-316 (2002).
16. Feng, Q. *et al*. Sequence and analysis of rice chromosome 4. *Nature* **420**, 316-320 (2002).
17. Rice Chromosome 10 Sequencing Consortium. In-depth view of structure, activity, and evolution of rice chromosome 10. *Science* **300**, 1566-1569 (2003).
18. *Arabidopsis* Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796-815 (2000).
19. Lowe, T. M. & Eddy, S. R. tRNA_scan SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res*. **25**, 955-964 (1997).
20. Griffiths-Jones S. The microRNA Registry. *Nucleic Acids Res.* **32**, D109-111 (2004).
21. Pearson, W. R. & Lipman, D. J. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA* **85**, 2444-2448 (1988).
22. Hiratsuka, J. *et al.* The complete sequence of the rice (*Oryza sativa*) chloroplast genome: Intermolecular recombination between distinct tRNA genes accounts for a major plastid DNA inversion during the evolution of the cereals. *Mol. Gen. Genet*. **217**, 185-194 (1989).
23. Notsu, Y. *et al.* The complete sequence of the rice (*Oryza sativa* L.) mitochondrial genome: frequent DNA sequence acquisition and loss during the evolution of flowering plants. *Mol. Gene. Genomics* **268**, 434-445 (2002)
24. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol*. **215**, 403-410 (1990).
25. Delcher, A. L. *et al.* Alignment of whole genomes. *Nucleic Acids Res*. **27**, 2369-2376 (1999).
26. Juretic, N., Bureau, T. E. & Bruskiewich, R. M. Transposable element annotation of the rice genome. *Bioinformatics* **20**,155-160 (2004).
27. Feschotte, C. & Wessler, S. R. *Mariner*-like transposases are widespread and diverse in flowering plants. *Proc. Natl. Acad. Sci. USA*. **99**, 280-285 (2002).
28. Inukai, T. & Sano, Y. Sequence rearrangement in the AT-rich minisatellite of the novel rice transposable element *Basho*. *Genome* **45** 493-502 (2002).
29. Iwamoto, M., Nagashima, H., Nagamine, T., Higo, H. & Higo, K. A Tourist element in the 5'-flanking region of the catalase gene *CatA* reveals evolutionary relationships among *Oryza* species with various genome types. *Mol. Gen. Genet*. **262**, 493-500 (1999).
30. Jiang, N. *et al*. Dasheng: a recently amplified nonautonomous long terminal repeat element that is a major component of pericentromeric regions in rice. *Genetics* **161**, 1293-1305 (2002).
31. Jiang, N. *et al.* An active DNA transposon family in rice. *Nature* **421**, 163-167 (2003).
32. Kikuchi, K., Terauchi, K., Wada, M. & Hirano, H. Y. The plant MITE *mPing* is mobilized in tissue culture. *Nature* **421**, 167-170 (2003).
33. Ma, J., Devos, K. M. & Bennetzen, J. L. Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice. *Genome Res.* **14**, 860-869 (2004).

34. McCarthy, E. M., Liu, J., Lizhi, G. & McDonald, J. F. Long terminal repeat retrotransposons of *Oryza sativa*. *Genome Biology* **3**, research 0053.0-0053.11 (2002).
35. Nakazaki, T. *et al.* Mobilization of a transposon in the rice genome. *Nature* **421**, 170-172 (2003).
36. Panaud, O. *et al*. Characterization of transposable elements in the genome of rice (*Oryza sativa* L.) using Representational Difference Analysis (RDA). *Mol. Genet. Genomics* **268**, 113-121 (2002).
37. Turcotte, K., Srinivasan, S. & Bureau, T. Survey of TEs from rice genomic sequences. *Plant J*. **25**, 169-179 (2001).
38. Yang, G, Dong, J., Chandrasekharan, M. B. & Hall, T. C. Kiddo, a new transposable element family closely associated with rice genes. *Mol. Genet. Genomics*. **266**, 417-424 (2001).
39. Yang, G. & Hall, T. C. MDM-1 and MDM-2: 2 Mutator-derived MITE families in rice. *J. Mol. Evol*. **56**, 255-264 (2003).
40. Zhang, X. *et al. P* instability factor: an active maize transposon system associated with the amplification of *Tourist*-like MITEs and a new superfamily of transposases. *Proc. Natl. Acad. Sci. USA*. **98**, 12572-12577 (2001).
41. Eddy, S. R. Profile hidden Markov models. *Bioinformatics* **14**, 755-763 (1998).
42. Petrov, D. A., Aminetzach, Y. T., Davis, J. C., Bensasson, D. & Hirsch, A. E. Size matters: non-LTR retrotransposable elements and ectopic recombination in *Drosophila. Mol. Biol. Evol*. **20**, 880-892 (2003).
43. Miyao, A., Yamazaki, M., & Hirochika, H. Systematic screening of mutants of rice by sequencing retrotransposon-insertion sites. *Plant Biotech*. **15**, 253-256 (1998).
44. Katagiri, S. *et al.* End sequencing and chromosomal *in silico* mapping of BAC clones derived from an *indica* rice cultivar, Kasalath. *Breeding Sci*. **54**, 273-279 (2004).
45. Temnykh, S. *et al.* Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): frequency, length variation, transposon associations, and genetic marker potential. *Genome Res*. **11**, 1441-52 (2001).
46. Harushima, Y. *et al.* A high-density rice genetic linkage map with 2275 markers using a single F2 population. *Genetics* **148**, 479-494 (1998).
47. Causse, M. A., *et al*. Saturated molecular map of the rice genome based on an interspecific backcross population. *Genetics* **138**,1251-1274 (1994).
48. McCouch, S.R., *et al*. Development and mapping of 2240 new SSR markers for rice (*Oryza sativa* L.). *DNA Res*. **9**, 199-207 (2002).
49. Yu, J. *et al.* The genomes of *Oryza sativa*: A history of duplications. *PLoS Biol.* **3**, e38, 1-16 (2005).
50. Feltus, F.A. *et al.* An SNP resource for rice genetics and breeding based on subspecies indica and japonica genome alignments. *Genome Res*. **14,** 1812-1819 (2004).
51. Shen, Y. J. *et al.* Development of genome-wide DNA polymorphism database for map-based cloning of rice genes. *Plant Physiol*. **135**, 1198-1205 (2004).
52. Han, B. & Xue, Y. Genome-wide intraspecific DNA-sequence variations in rice. *Curr. Opin. Plant Biol*. **6**, 134-138 (2003).

Legends for Supplementary Figures

Supplementary Figure 1. Rice gene models that align with cereal EST sequences. This data may be biased by the depth of species-specific cDNA libraries and does not reflect evolutionary distances.

Supplementary Figure 2. The density of expressed genes on the twelve rice chromosomes. The 12 chromosomes (in green) are depicted from 1 to 12 (top to bottom) with the short arms to the left. The frequency of gene models with EST or full-length cDNA hits in 100 kb windows are plotted in blue. Centromeres (red boxes) and physical gaps (white spaces) are indicated on the chromosomes.

Supplementary Figure 3. Functional classification of predicted proteins and protein domains.
a. Biological process (proteins)
b. Molecular function (domains)

Supplementary Figure 4. Comparison of the predicted rice proteome with other model species. A total of 37,544 rice proteins were searched against the protein datasets of *Arabidopsis*, *S. cerevisiae*, *C. elegans*, *Drosophila*, human, *Synechocystis* and *E. coli* using BLASTP and the scores were binned into six categories.

Supplementary Figure 5. Distribution of *Tos17* insertions on the twelve rice chromosomes. The frequency of aligned flanking sequences in a 100 kb window is plotted. Positions of centromeric repeats (CentO) are indicated. The chromosomes are oriented from short arm (left) to long arm (right).

Supplementary Figure 6. Array sizes of tandemly repeated genes.

Supplementary Figure 7. The density of tRNA genes on the twelve rice chromosomes. The frequency of tRNA genes in a 100 kb window is plotted. The chromosomes are oriented from short arm (left) to long arm (right).

Supplementary Figure 8. Distribution of organellar inserts on the Nipponbare pseudomolecules. Chloroplast sequences (green) appear to the left of each chromosome, mitochondrial sequences (red) to the right, and sequences that have identity with both organelles (black) are shown on both sides of the chromosome. The thickness of the lines is roughly proportional to the length of the insert although a series of small inserts may appear as one thick line. The thinnest bars represent inserts 100 to 60,000 bp long.

Supplementary Figure 9. Frequency of the ten most frequent SSR motif families in the rice genome.

Supplementary Figure 10. Comparison of the 93-11 and Syngenta assemblies with a portion of IRGSP chromosome 1S. The BGI assembled and mapped Syngenta contigs (AACV01000001-AACV01000070) and the 93-11 contigs (AAAA02000001-AAAA02000093) were compared by BLAST analysis against the IRGSP chromosome 1 pseudomolecule.

Supplementary Table1 **Libraries used in construction of physical maps**

Library name	Clone ID	Vector	Enzyme	Ave. insert size (kb)	No. of clones	Finger-printed	End-sequenced	Clones sequenced
RGP YAC[a]	Y	YAC	EcoRI, NotI	350	6932	-	-	-
RGP PAC[b]	P	PAC	Sau3A1	112	71040	Partly	Partly	907
RGP BAC[c]	B	BAC	MboI	124	48960	Partly	Partly	222
CUGI BAC[d]	OSJNBa	BAC	HindIII	129	36864	Fully	Fully	1081
CUGI BAC	OSJNBb	BAC	EcoRI	119	55296	Fully	Fully	553
Monsanto BAC[e]	OJ	BAC	HindIII	-	3416	Fully	Fully	638
AGI fosmid[f]	OSJNOa	Fosmid	Sheared DNA	41	110592	-	-	18
CUGI plasmid[g]	OSJNPb	Plasmid	HaeIII	10	165888	-	-	-
CUGI plasmid	OSJNPc	Plasmid	Sau3AI	10	138240	-	-	-
Others[h]	OSJNA, OJA	-	-	-	-	-	-	34
Total					630296			3453

[a]Rice Genome Research Program YAC: Saji, S., *et al*. *Genome* **44**, 32-37 (2001)

[b]Rice Genome Research Program PAC, Baba, T.,*et al*. *Bull. Natl. Inst. Agrobiol. Resour. (Japan)* **14**, 41-51 (2000)

[c]Rice Genome Research Program BAC, Wu, J., *et al*. *Plant J.* **36**, 720-730 (2003)

[d]Clemson University Genomics Institute BAC: Chen, M., *et al*. *Plant Cell* **14**, 537-545 (2002)

[e]Monsanto BAC: Barry, G. F.,*Plant Physiol.* **125**, 1164-1165 (2001)

[f]Arizona Genomics Institute fosmid: http://www.genome.arizona.edu/orders/direct.html?library=OSJNOa

[g]Clemson University Genomics Institute plasmid: Yang, T. J.,*et al*. *Theor. Appl. Genet*. **107**, 652-660 (2003)

[h]OSJNA and OJA: Artificial gap-filling clones.

Supplementary Table 2 **Sequence quality based on overlapping sequences**

Chr	Overlapping clones		Overlap sequence (bp)	Base substitutions (bp)	Insertion / deletion (bp)
1	OSJNBa0049B20(TIGR)	OSJNBa0004G10(RGP)	117850	2	8
1	OSJNBa0049B20(TIGR)	P0034C11(RGP)	136361	1	8
1	OSJNBa0048I01(KRGP)	P408G07(RGP)	185780	7	32
1	OSJNBa0048I01(KRGP)	B1099D03(RGP)	58337	0	7
2	OSJNBa0049B20(CSHL)	OJ1111_C07(RGP)	150046	0	40
6	OJ1540_H01(TIGR)	P0481E08(RGP)	57158	0	9
6	OJ1540_H01(TIGR)	P0541C02(RGP)	88194	0	0
7	OSJNBb0024A20(ACWW)	OSJNBa0072I06(RGP)	19465	0	0
7	OSJNBb0024A20(ACWW)	OSJNBb0018L13(RGP)	129833	0	20
10	OJ1004_F02(TIGR)	OSJNBa0014J14(ACWW)	85048	0	0
10	OSJNBa0093I09(ACWW)	OSJNBa0073L20(TIGR)	82117	0	10
11	OSJNBa0052C03(RGIR)	OSJNBa0052C16(TIGR)	45294	0	0
11	OSJNBa0094P07(TIGR)	OSJNBb0088N01(PGIR)	71138	0	0
11	OSJNBa0025K19(Genoscope)	OSJNBb0004B05(PGIR)	21264	0	0
Totals			1247885	10	134
Accuracy based on base pair discrepancies (%)					99.9992
Accuracy based on both substitutions and insertions/deletions (%)					99.9885

Supplementary Table 3 **CentO (155 bp satellite DNA) units within chromosome pseudomolecules**

Chr	Physical map	CentO sequence location[a]	Total units	Total amount (bp)	Identity[b] (%)
1	partial	16682539-17130082	1055	163525	82.3-97.6
2	partial	13570041-13857135	1297	201035	84.0-97.9
3	partial	19346905-19452749	158	24490	80.2-97.2
4	complete	9808276-9933189	355	55025	82.2-96.7
5	complete	12357944-12421998	325	50375	84.3-96.5
6	partial	15266486-15272427	39	6045	81.2-97.4
7	partial	11992162-12227761	578	89590	84.3-97.8
8	complete	12913343-13833051	443	68665	81.9-99.3
9	partial	2697206-2927046	776	120280	85.3-95.0
10	partial	7701335-8609236	7	1085	83.0-92.9
11	partial	11939274-11939374	-	-	-
12	partial	11771323-12117142	814	126170	79.4-97.8
All			5847	906285	

[a]Pseudomolecule coordinates.

[b]The 155-bp consensus CentO seuqence analyzed from the centromeric region of chromosome 8 was used for Blast analysis.

Supplementary Table 4 **Chromosomal distribution of gene models**

Chromosome	Length (bp)	Predicted models	Gene Density (kbp/gene)
1	43260640	4856	8.9
2	35954074	3964	9.1
3	36189985	4159	8.7
4	35489479	3400	10.4
5	29733216	2956	10.1
6	30731386	3079	10
7	29643843	3044	9.7
8	28434680	2708	10.5
9	22692709	2175	10.4
10	22683701	2185	10.4
11	28357783	2650	10.7
12	27561960	2368	11.6
Total	370733456	37544	9.9

Supplementary Table 5 **Statistics for the predicted genes in the rice pseudomolecules and a comparison with *Arabidopsis thaliana***

	Rice pseudomolecules	*Arabidopsis* genome[a]
Length (bp)	370,733,456	115,409,949
Predicted genes		
Number	37,544	25,498
Gene density (kb per gene)	9.9	4.5
Average gene length (bp)	2,699	1,992
Exons		
Number	175,203	132,982
Total length (bp)	44,492,676	33,249,250
Average per gene	4.7	5.2
Average size (bp)	254	250
Introns		
Number	137,659	107,484
Total length (bp)	56,841,388	18,055,421
Average per gene	3.7	4.2
Average size (bp)	413	168
Base composition (GC %)		
Exon	54.2	44.1
Intron	38.3	32.7
Intergenic	42.9	
Gene	45.3	
Genome	43.6	34.7

[a]Arabidopsis Genome Initiative. *Nature* **408**, 796-815 (2000).

Supplementary Table 6 **Coverage of FGENESH models with FL-cDNAs**

Cutoff length (%)	Alignments	FGENESH models	FLcDNAs
25	30,253	17,016	25,636
50	23,581	14,907	22,046
75	17,690	11,534	16,719
80	16,400	10,806	15,513

FGENESH models (37,544) were searched using BLASTN against the collection of 32,767 FL-cDNAs. The alignments were parsed using 95% identity over variable length cutoffs.

Supplementary Table 7 **The 50 most frequent domains detected by Interpro**

	IPRid	Description	Gene models
1	IPR011009	Protein kinase-like	1425
2	IPR000719	Protein kinase	1366
3	IPR002290	Serine/threonine protein kinase	1286
4	IPR001245	Tyrosine protein kinase	1264
5	IPR008271	Serine/threonine protein kinase, active site	1075
6	IPR001611	Leucine-rich repeat	837
7	IPR001810	Cyclin-like F-box	620
8	IPR008941	TPR-like	572
9	IPR007090	Leucine-rich repeat, plant specific	431
10	IPR009057	Homeodomain-like	425
11	IPR002885	PPR repeat	423
12	IPR001841	Zn-finger, RING	401
13	IPR002182	NB-ARC	396
14	IPR000767	Disease resistance protein	361
15	IPR008938	ARM repeat fold	337
16	IPR001128	Cytochrome P450	331
17	IPR001005	Myb, DNA-binding	328
18	IPR003591	Leucine-rich repeat, typical subtype	321
19	IPR008940	Protein prenyltransferase	305
20	IPR000504	RNA-binding region RNP-1 (RNA recognition motif)	279
21	IPR002401	E-class P450, group I	274
22	IPR000345	Cytochrome c heme-binding site	262
23	IPR009007	Peptidase aspartic	242
24	IPR003593	AAA ATPase	240
25	IPR001680	G-protein beta WD-40 repeat	233
26	IPR002048	Calcium-binding EF-hand	218
27	IPR002110	Ankyrin	214
28	IPR000379	Esterase/lipase/thioesterase	210
29	IPR011046	WD40-like	188
30	IPR010983	EF-Hand-like	187
31	IPR007087	Zn-finger, C2H2 type	174
32	IPR001092	Basic helix-loop-helix dimerisation region bHLH	162
33	IPR001471	Pathogenesis-related transcriptional factor and ERF	157
34	IPR002016	Haem peroxidase, plant/fungal/bacterial	154
35	IPR001878	Zn-finger, CCHC type	154
36	IPR003612	Plant lipid transfer/seed storage/trypsin-alpha amylase inhi	150
37	IPR002213	UDP-glucuronosyl/UDP-glucosyltransferase	146
38	IPR008974	TRAF-like	139
39	IPR010255	Haem peroxidase	139
40	IPR001440	TPR repeat	139
41	IPR008985	Concanavalin A-like lectin/glucanase	137
42	IPR001687	ATP/GTP-binding site motif A (P-loop)	137
43	IPR003439	ABC transporter	129
44	IPR000210	BTB/POZ	127
45	IPR008994	Nucleic acid-binding OB-fold	124
46	IPR000823	Plant peroxidase	123
47	IPR011011	FYVE/PHD zinc finger	122
48	IPR001410	DEAD/DEAH box helicase	120
49	IPR001480	Curculin-like (mannose-binding) lectin	119
50	IPR003441	No apical meristem (NAM) protein	117

Supplementary Table 8 **Cereal-specific proteins**

Protein	Number
Abscisic stress ripening protein	4
Chitinase precursor	3
Citrate binding protein precursor	1
Endonuclease	1
Glucan 1,3-beta-glucosidase precursor	3
Heterogenous nuclear ribonucleoprotein	1
Jasomate-induced protein	4
Mannosyltransferase	1
Pathogenesis-related protein PR-10a	5
Phytosulfokines precursor	1
Prolamin	31
Proteinase inhibitor	10
Queuine tRNA-ribosyltransferase	2
Ribosome-inactivating protein	1
SAM-dependent methyltransferase	1
Seed allergen	5
Starch branching enzyme	1
Wound-induced protease inhibitor	1

Supplementary Table 9 **Motifs in tandemly repeated gene families**

IPR Number[a]	Motif Description	Gene models with motif/domain	Gene models in tandem arrays
IPR011009	Protein kinase-like	125	134
IPR002290	Serine/threonine protein kinase	122	134
IPR001245	Tyrosine protein kinase	121	134
IPR008271	Serine/threonine protein kinase, active site	107	134
IPR011009	Protein kinase-like	92	109
IPR001245	Tyrosine protein kinase	91	109
IPR002290	Serine/threonine protein kinase	91	109
IPR000719	Protein kinase	91	109
IPR008271	Serine/threonine protein kinase, active site	84	109
IPR000719	Protein kinase	85	100
IPR011009	Protein kinase-like	85	100
IPR002290	Serine/threonine protein kinase	81	100
IPR001245	Tyrosine protein kinase	80	100
IPR008271	Serine/threonine protein kinase, active site	70	100
IPR002182	NB-ARC	77	91
IPR000767	Disease resistance protein	66	91
IPR001611	Leucine-rich repeat	64	91
IPR001611	Leucine-rich repeat	57	73
IPR007090	Leucine-rich repeat, plant specific	55	73
IPR011009	Protein kinase-like	53	73
IPR002290	Serine/threonine protein kinase	53	73
IPR001245	Tyrosine protein kinase	53	73
IPR000719	Protein kinase	53	73
IPR008271	Serine/threonine protein kinase, active site	45	73
IPR003591	Leucine-rich repeat, typical subtype	44	73
IPR011009	Protein kinase-like	56	69
IPR000719	Protein kinase	56	69
IPR001245	Tyrosine protein kinase	55	69
IPR002290	Serine/threonine protein kinase	54	69
IPR008271	Serine/threonine protein kinase, active site	49	69
IPR011009	Protein kinase-like	55	57
IPR000719	Protein kinase	54	57
IPR002290	Serine/threonine protein kinase	53	57
IPR001245	Tyrosine protein kinase	53	57
IPR008271	Serine/threonine protein kinase, active site	46	57
IPR002885	PPR repeat	51	51
IPR008941	TPR-like	43	51
IPR008940	Protein prenyltransferase	37	51
IPR011009	Protein kinase-like	50	50
IPR002290	Serine/threonine protein kinase	50	50
IPR001245	Tyrosine protein kinase	50	50
IPR000719	Protein kinase	50	50
IPR000719	Protein kinase	45	50
IPR011009	Protein kinase-like	45	50
IPR002290	Serine/threonine protein kinase	45	50
IPR001245	Tyrosine protein kinase	45	50
IPR008271	Serine/threonine protein kinase, active site	37	50
IPR000719	Protein kinase	41	45
IPR002290	Serine/threonine protein kinase	41	45
IPR011009	Protein kinase-like	41	45
IPR001245	Tyrosine protein kinase	41	45
IPR002885	PPR repeat	44	44
IPR008941	TPR-like	36	44
IPR008940	Protein prenyltransferase	29	44
IPR001245	Tyrosine protein kinase	42	43

IPR000719	Protein kinase	42	43
IPR011009	Protein kinase-like	42	43
IPR002290	Serine/threonine protein kinase	42	43
IPR008271	Serine/threonine protein kinase, active site	34	43
IPR008974	TRAF-like	40	42
IPR000210	BTB/POZ	40	42
IPR002083	MATH	35	42
IPR002182	NB-ARC	40	41
IPR000767	Disease resistance protein	39	41
IPR001611	Leucine-rich repeat	31	41
IPR011009	Protein kinase-like	37	37
IPR001245	Tyrosine protein kinase	37	37
IPR000719	Protein kinase	37	37
IPR002290	Serine/threonine protein kinase	37	37
IPR008271	Serine/threonine protein kinase, active site	33	37
IPR001128	Cytochrome P450	36	36
IPR002401	E-class P450, group I	33	36
IPR001245	Tyrosine protein kinase	31	35
IPR002290	Serine/threonine protein kinase	31	35
IPR011009	Protein kinase-like	31	35
IPR000719	Protein kinase	31	35
IPR008271	Serine/threonine protein kinase, active site	25	35
IPR001611	Leucine-rich repeat	24	35
IPR007090	Leucine-rich repeat, plant specific	24	35
NO IPR			32
IPR001128	Cytochrome P450	32	32
IPR002401	E-class P450, group I	29	32
IPR002182	NB-ARC	29	31
IPR000767	Disease resistance protein	28	31
IPR001611	Leucine-rich repeat	25	31
IPR002885	PPR repeat	30	30
IPR008941	TPR-like	26	30
IPR008940	Protein prenyltransferase	19	30
IPR002182	NB-ARC	27	29
IPR000767	Disease resistance protein	26	29
IPR004045	Glutathione S-transferase, N-terminal	28	28
IPR010987	Glutathione S-transferase, C-terminal-like	28	28
IPR004046	Glutathione S-transferase, C-terminal	27	28
IPR002952	Eggshell protein	23	28
IPR001245	Tyrosine protein kinase	27	27
IPR000719	Protein kinase	27	27
IPR002290	Serine/threonine protein kinase	27	27
IPR011009	Protein kinase-like	27	27
IPR008271	Serine/threonine protein kinase, active site	19	27
IPR011009	Protein kinase-like	23	27
IPR002290	Serine/threonine protein kinase	23	27
IPR000719	Protein kinase	23	27
IPR001245	Tyrosine protein kinase	23	27
IPR008271	Serine/threonine protein kinase, active site	20	27
IPR007658	Protein of unknown function DUF594	20	26
IPR002885	PPR repeat	26	26
IPR008941	TPR-like	21	26
IPR008940	Protein prenyltransferase	17	26
NO IPR			25
IPR002182	NB-ARC	21	25
IPR000767	Disease resistance protein	20	25
IPR001283	Allergen V5/Tpx-1 related	23	23
NO IPR			23

IPR004320	Arabidopsis conserved protein	20	23
IPR002885	PPR repeat	23	23
IPR008941	TPR-like	22	23
IPR008940	Protein prenyltransferase	15	23
IPR001810	Cyclin-like F-box	20	23
IPR002182	NB-ARC	22	22
IPR000767	Disease resistance protein	22	22
IPR001611	Leucine-rich repeat	20	22
IPR002213	UDP-glucuronosyl/UDP-glucosyltransferase	16	22
IPR002885	PPR repeat	22	22
IPR008941	TPR-like	21	22
IPR002885	PPR repeat	21	21
IPR008941	TPR-like	20	21
IPR008940	Protein prenyltransferase	14	21
NO IPR			21
IPR005299	SAM dependent carboxyl methyltransferase	18	21
IPR002213	UDP-glucuronosyl/UDP-glucosyltransferase	18	20
IPR002401	E-class P450, group I	20	20
IPR001128	Cytochrome P450	20	20
IPR000767	Disease resistance protein	20	20
IPR002182	NB-ARC	19	20
IPR001611	Leucine-rich repeat	18	20
NO IPR			19
IPR002182	NB-ARC	19	19
IPR000767	Disease resistance protein	18	19
IPR001611	Leucine-rich repeat	18	19
NO IPR			19
IPR000767	Disease resistance protein	18	18
IPR002182	NB-ARC	17	18
IPR001611	Leucine-rich repeat	15	18
NO IPR			18
NO IPR			18
IPR001611	Leucine-rich repeat	18	18
IPR007090	Leucine-rich repeat, plant specific	18	18
IPR003591	Leucine-rich repeat, typical subtype	17	18
IPR002016	Haem peroxidase, plant/fungal/bacterial	17	17
IPR010255	Haem peroxidase	17	17
IPR000823	Plant peroxidase	17	17
NO IPR			17
IPR003612	Plant lipid transfer/seed storage/trypsin-alpha amylase inhibitor	13	17
IPR000719	Protein kinase	17	17
IPR008271	Serine/threonine protein kinase, active site	17	17
IPR011009	Protein kinase-like	17	17
IPR001245	Tyrosine protein kinase	17	17
IPR002290	Serine/threonine protein kinase	17	17
IPR002885	PPR repeat	17	17
IPR008941	TPR-like	15	17
IPR008940	Protein prenyltransferase	11	17
IPR009007	Peptidase aspartic	16	17
IPR002110	Ankyrin	17	17
IPR001810	Cyclin-like F-box	13	17
IPR001810	Cyclin-like F-box	14	17
IPR011009	Protein kinase-like	17	17
IPR000719	Protein kinase	17	17
IPR002290	Serine/threonine protein kinase	16	17
IPR001245	Tyrosine protein kinase	16	17
IPR008271	Serine/threonine protein kinase, active site	16	17
IPR000490	Glycoside hydrolase, family 17	16	16

IPR009007	Peptidase aspartic	11	16
NO IPR			16
IPR002213	UDP-glucuronosyl/UDP-glucosyltransferase	12	16
IPR002885	PPR repeat	16	16
IPR008940	Protein prenyltransferase	12	16
IPR008941	TPR-like	12	16
IPR002213	UDP-glucuronosyl/UDP-glucosyltransferase	15	16
IPR011009	Protein kinase-like	16	16
IPR002290	Serine/threonine protein kinase	16	16
IPR001245	Tyrosine protein kinase	16	16
IPR000719	Protein kinase	16	16
IPR008271	Serine/threonine protein kinase, active site	14	16
IPR000210	BTB/POZ	16	16
IPR008974	TRAF-like	15	16
IPR002083	MATH	12	16
IPR001128	Cytochrome P450	15	15
IPR002401	E-class P450, group I	12	15
IPR000719	Protein kinase	15	15
IPR001245	Tyrosine protein kinase	15	15
IPR011009	Protein kinase-like	15	15
IPR002290	Serine/threonine protein kinase	15	15
IPR008271	Serine/threonine protein kinase, active site	11	15
IPR002401	E-class P450, group I	14	15
IPR001128	Cytochrome P450	14	15
IPR005123	2OG-Fe(II) oxygenase superfamily	15	15
IPR004253	Protein of unknown function DUF231	14	15
IPR001128	Cytochrome P450	15	15
IPR002401	E-class P450, group I	10	15
IPR002885	PPR repeat	15	15
IPR008941	TPR-like	14	15
IPR008940	Protein prenyltransferase	11	15
IPR000379	Esterase/lipase/thioesterase	14	15
NO IPR			15
NO IPR			15
IPR002885	PPR repeat	15	15
IPR008941	TPR-like	11	15
IPR008940	Protein prenyltransferase	10	15
IPR002110	Ankyrin	14	15
IPR001611	Leucine-rich repeat	14	14
IPR007090	Leucine-rich repeat, plant specific	14	14
IPR003591	Leucine-rich repeat, typical subtype	12	14
IPR002885	PPR repeat	14	14
IPR008940	Protein prenyltransferase	12	14
IPR008941	TPR-like	11	14
IPR002213	UDP-glucuronosyl/UDP-glucosyltransferase	13	14
IPR009007	Peptidase aspartic	13	14
IPR001810	Cyclin-like F-box	11	14
IPR003676	Auxin responsive SAUR protein	14	14
IPR003480	Transferase	13	14
IPR001878	Zn-finger, CCHC type	9	14
IPR001087	Lipolytic enzyme, G-D-S-L	11	13
IPR002213	UDP-glucuronosyl/UDP-glucosyltransferase	11	13
IPR001128	Cytochrome P450	13	13
IPR002401	E-class P450, group I	13	13
IPR001245	Tyrosine protein kinase	13	13
IPR000719	Protein kinase	13	13
IPR011009	Protein kinase-like	13	13
IPR002290	Serine/threonine protein kinase	13	13

IPR008271	Serine/threonine protein kinase, active site	9	13
IPR010255	Haem peroxidase	13	13
IPR002016	Haem peroxidase, plant/fungal/bacterial	13	13
IPR000823	Plant peroxidase	13	13
IPR001810	Cyclin-like F-box	11	13
IPR010616	Protein of unknown function DUF1210	13	13
IPR008941	TPR-like	13	13
IPR002885	PPR repeat	13	13
IPR007118	Expansin/Lol pI	12	12
IPR007112	Expansin 45, endoglucanase-like	12	12
IPR009009	Barwin-related endoglucanase	12	12
IPR007117	Pollen allergen/expansin, C-terminal	12	12
IPR001245	Tyrosine protein kinase	12	12
IPR002290	Serine/threonine protein kinase	12	12
IPR000719	Protein kinase	12	12
IPR011009	Protein kinase-like	12	12
IPR008271	Serine/threonine protein kinase, active site	11	12
IPR005630	Terpene synthase, metal-binding	12	12
IPR001906	Terpene synthase-like	12	12
IPR008949	Terpenoid synthase	12	12
IPR008930	Terpenoid cylases/protein prenyltransferase alpha-alpha toroid	11	12
IPR000767	Disease resistance protein	12	12
IPR002182	NB-ARC	11	12
IPR001611	Leucine-rich repeat	9	12
IPR011009	Protein kinase-like	12	12
IPR008271	Serine/threonine protein kinase, active site	12	12
IPR000719	Protein kinase	12	12
IPR002290	Serine/threonine protein kinase	12	12
IPR001245	Tyrosine protein kinase	12	12
IPR002182	NB-ARC	12	12
IPR000767	Disease resistance protein	10	12
IPR001611	Leucine-rich repeat	8	12
IPR009007	Peptidase aspartic	8	12
IPR002213	UDP-glucuronosyl/UDP-glucosyltransferase	10	12
IPR002347	Glucose/ribitol dehydrogenase	12	12
IPR002198	Short-chain dehydrogenase/reductase SDR	12	12
NO IPR			12
IPR007113	Cupin region	12	12
IPR011051	RmlC-like cupin	12	12
IPR006045	Cupin	12	12
IPR001929	Germin	12	12
IPR001810	Cyclin-like F-box	11	12
IPR000668	Peptidase C1A, papain	12	12
IPR000169	Peptidase, eukaryotic cysteine peptidase active site	11	12
IPR004265	Plant disease resistance response protein	11	12
IPR001810	Cyclin-like F-box	9	12
NO IPR			11
IPR006041	Pollen Ole e 1 allergen and extensin	8	11
IPR001810	Cyclin-like F-box	9	11
IPR000109	TGF-beta receptor, type I/II extracellular region	11	11
IPR001128	Cytochrome P450	11	11
IPR002401	E-class P450, group I	10	11
IPR001087	Lipolytic enzyme, G-D-S-L	11	11
IPR002213	UDP-glucuronosyl/UDP-glucosyltransferase	10	11
IPR001810	Cyclin-like F-box	9	11
IPR001810	Cyclin-like F-box	10	11
NO IPR			11
IPR000823	Plant peroxidase	11	11

IPR002016	Haem peroxidase, plant/fungal/bacterial	11	11
IPR010255	Haem peroxidase	11	11
IPR008938	ARM repeat fold	10	11
IPR000379	Esterase/lipase/thioesterase	11	11
IPR001810	Cyclin-like F-box	8	11
NO IPR			11
IPR000767	Disease resistance protein	11	11
IPR002182	NB-ARC	11	11
IPR001611	Leucine-rich repeat	9	11
IPR001611	Leucine-rich repeat	11	11
IPR007090	Leucine-rich repeat, plant specific	11	11
IPR002290	Serine/threonine protein kinase	10	11
IPR011009	Protein kinase-like	10	11
IPR003591	Leucine-rich repeat, typical subtype	10	11
IPR000719	Protein kinase	10	11
IPR001245	Tyrosine protein kinase	10	11
IPR008271	Serine/threonine protein kinase, active site	8	11
IPR009007	Peptidase aspartic	11	11
IPR001461	Peptidase A1, pepsin	10	11
IPR001223	Glycoside hydrolase, family 18	11	11
IPR000209	Peptidase S8 and S53, subtilisin, kexin, sedolisin	10	10
IPR003137	Protease-associated PA	10	10
IPR009020	Proteinase inhibitor, propeptide	9	10
IPR010259	Proteinase inhibitor I9, subtilisin propeptide	9	10
NO IPR			10
IPR001841	Zn-finger, RING	10	10
NO IPR			10
IPR001128	Cytochrome P450	10	10
IPR002401	E-class P450, group I	10	10
IPR005829	Sugar transporter superfamily	9	10
IPR003663	Sugar transporter	9	10
IPR007114	Major facilitator superfamily	9	10
IPR005828	General substrate transporter	9	10
IPR002182	NB-ARC	10	10
IPR000767	Disease resistance protein	10	10
IPR001611	Leucine-rich repeat	7	10
IPR000209	Peptidase S8 and S53, subtilisin, kexin, sedolisin	7	10
IPR003137	Protease-associated PA	7	10
IPR002213	UDP-glucuronosyl/UDP-glucosyltransferase	9	10
IPR001360	Glycoside hydrolase, family 1	10	10
IPR009003	Peptidase, trypsin-like serine and cysteine proteases	9	10
IPR001478	PDZ/DHR/GLGF	8	10
IPR001320	Ionotropic glutamate receptor	10	10
IPR001311	Solute-binding protein/glutamate receptor	8	10
IPR011009	Protein kinase-like	10	10
IPR001480	Curculin-like (mannose-binding) lectin	10	10
IPR000719	Protein kinase	10	10
IPR008271	Serine/threonine protein kinase, active site	9	10
IPR001245	Tyrosine protein kinase	9	10
IPR002290	Serine/threonine protein kinase	9	10
IPR003480	Transferase	10	10
NO IPR			10
IPR011009	Protein kinase-like	10	10
IPR002182	NB-ARC	8	10
IPR011009	Protein kinase-like	10	10
IPR003612	Plant lipid transfer/seed storage/trypsin-alpha amylase inhibitor	10	10
NO IPR			10
IPR000210	BTB/POZ	10	10
IPR008974	TRAF-like	10	10
IPR002083	MATH	10	10
IPR001938	Thaumatin, pathogenesis-related	10	10

[a]Domains detected in Interpro for the gene clusters of ten or more members.

Supplementary Table 10 **MicroRNAs in the *Oryza* genome**

Chr	*Oryza* miRNA[a]	Other miRNA[b]	Total
1	17	1	18
2	21	2	23
3	11	1	12
4	18	1	19
5	7	5	12
6	11	4	15
7	6	4	10
8	15	4	19
9	7	0	7
10	5	1	6
11	5	2	7
12	6	4	10
Total	129	29	158

[a]Predicted homologues of *Arabidopsis* miRNAs from the Rfam database.

[b]Homologues of experimentally validated miRNAs of other species excluding *Arabidopsis.*

Supplementary Table 11 **snoRNA and splicesomal RNA genes in the *Oryza* genome**

Chr	snoRNA genes	Splicesomal RNA genes
1	13	6
2	17	25
3	64	13
4	12	9
5	19	1
6	16	11
7	22	10
8	14	6
9	5	0
10	14	4
11	7	5
12	12	3
Total	215	93

Supplementary Table 12 **Organellar sequences in the Nipponbare chromosomes**

A **Chloroplast inserts**

Chr	MUMmer with high stringency			BLAST with medium stringency		
	Inserts (No.)	Total length (bp)	% of chromosome	Inserts (No.)	Total length (bp)	% of chomosome
1	71	78842	0.183	67	100307	0.233
2	43	67793	0.190	42	76179	0.213
3	63	38121	0.112	61	49043	0.145
4	47	86912	0.252	40	128137	0.372
5	24	37594	0.138	23	44081	0.162
6	35	45703	0.147	34	54531	0.175
7	25	30319	0.106	24	35699	0.124
8	22	51943	0.184	21	57576	0.204
9	21	10229	0.047	19	20515	0.094
10	20	165771	0.739	14	178504	0.796
11	29	10056	0.041	29	17127	0.007
12	53	79803	0.298	47	91214	0.341
Totals	**453**	**703086**	**0.196**	**421**	**852913**	**0.238**
Ave. % id.		**98.68**			**93.12**	
Genome eq.		**5.22**			**6.34**	

B **Mitochondrial inserts**

Chr	MUMmer with high stringency			BLAST with medium stringency		
	Inserts (No.)	Total length (bp)	% of chromosome	Inserts (No.)	Total length (bp)	% of chromosome
1	166	72223	0.168	197	78260	0.182
2	49	18901	0.053	57	19305	0.054
3	57	114731	0.338	92	119718	0.353
4	68	50294	0.146	94	53267	0.155
5	28	13809	0.051	32	13068	0.048
6	61	44952	0.144	79	44253	0.142
7	31	8687	0.03	30	8117	0.028
8	31	13306	0.047	44	15314	0.054
9	33	11174	0.051	41	17295	0.079
10	61	41487	0.185	85	41347	0.184
11	36	9320	0.038	50	10507	0.043
12	288	231573	0.865	390	269642	1.007
Totals	**909**	**630457**	**0.176**	**1191**	**690093**	**0.193**
Ave. % id.		**95.97**			**98.02**	
Genome eq.		**1.29**			**1.41**	

Supplementary Table 13 **Large organellar inserts in the Nipponbare genome**

A **Chloroplast inserts**

Chr	Start on chromo-some	Stop on chromo-some	Chromo-somal length (bp)	Start on ct	Stop on ct	Ct length (bp)	Identity (%)	Strand
10	10180595	10311746	131152	117385	113606	130747	98.74	D/R
8	9255091	9290202	35112	58789	93858	35070	99.34	R
10	19671024	19704027	33004	114484	125520	33385	99.22	R
6	23583508	23613061	29554	49883	79423	29541	98.87	R
4	8775246	8795361	20116	10330	30422	20093	99.14	D
4	8748156	8767894	19739	80045	99713	19669	98.49	D
12	5491125	5510781	19657	98906	118637	19732	98.58	D
2	14407382	14426139	18758	6890	25609	18720	98.57	R
4	8691010	8706795	15786	112456	128207	15752	99.18	R
4	8795362	8809793	14432	86406	101147	14742	98.25	D
7	13416762	13431163	14402	120124	134525	14402	99.76	D
12	5510782	5524143	13362	58215	71624	13410	98.61	D
4	8724611	8735889	11279	70435	81665	11231	98.94	R
Total bp in 13 inserts			376353		Mean % identity		98.88	

B **Mitochondrial inserts**

Chr	Start on chromo-some	Stop on chromo-some	Chromo-somal length (bp)	Start on mt	Stop on mt	Mt length (bp)	Identity (%)	Strand
12	19839260	19879661	40402	51833	92236	40404	99.71	R
12	19975413	19993653	18241	255406	273646	18241	99.85	R
12	19935657	19951039	15383	190687	206049	15363	99.48	D
3	22618288	22632207	13920	37915	51834	13920	99.09	R
12	19880779	19892524	11746	40088	51834	11747	98.8	R
12	19953387	19963860	10474	212838	223312	10475	99.51	D
6	12434844	12444862	10019	41292	51269	9978	95.62	D
Total bp in 7 inserts			120185		Mean % identity		99.18	

Supplementary Table 14 **Effect of average length on transposable element distribution patterns**

Transposable element family	Average length (bp)	Relative centromeric abundance[a]	Correlation with recombinationb	Correlation with gene density[b]
*h*AT	1554	0.84*	0.014	0.021
CACTA	1062	1.15**	-0.117**	-0.251**
Dasheng	1503	2.23**	-0.334**	-0.429**
LINEs	505	0.55**	0.265**	0.177**
solo LTRs	919	2.13**	-0.416**	-0.472**
other class II	154	0.68**	0.236**	0.152**
IS*630*/Tc1/*mariner*	139	0.56**	0.411**	0.320**
IS*256/Mutator*	1999	0.75**	0.122**	0.095**
SINEs	118	0.70**	0.139**	0.112**
IS*5/Tourist*	229	0.64**	0.346**	0.391**
other TEs	295	0.69**	0.208**	0.178**
TRIM	224	0.85	-0.016	-0.097**
Ty1/*copia*	1536	1.36**	-0.222**	-0.352**
Ty3/*gypsy*	1878	1.85**	-0.268**	-0.515**
genes		0.73**	0.355**	
exons		0.48**	0.376**	

[a]Abundance in centromeric and pericentromeric regions relative to random expectation. Values greater than 1 represent over-representation in centromeric and pericentromeric regions, while values less than 1 are under-representated. Significance evaluated using the chi square statistic with 1 degree of freedom.

[b]Spearman rank correlation coefficients between element abundance and correlation with recombination rate and gene density. Significance: *, $p<0.01$; **, $p<0.001$.

Supplementary Table 15 **Comparison of mapped Kasalath BAC-end sequences with the Nipponbare pseudomolecules**

Chr	Mapped clones	SNPs (bp)	Small InDels (bp)	Total alignment (bp)	SNP rate (%)
1	1830	10162	4028	1706473	0.60
2	1524	7562	3051	1427624	0.53
3	1686	8902	3514	1553638	0.57
4	1110	5923	2005	1019687	0.58
5	998	6813	2376	916803	0.74
6	1128	7452	2705	1037609	0.72
7	1002	6608	2327	925108	0.71
8	1034	6296	2399	965522	0.65
9	706	4973	1717	642492	0.77
10	821	5219	1770	762127	0.68
11	773	5212	1805	717266	0.73
12	704	5005	1701	644751	0.78
Total	**13316**	**80127**	**29398**	**12319100**	**0.65**

Supplementary Table 16 **Pattern of nucleotide substitution between Nipponbare and Kasalath**

Pattern	Chromosome												Total substitutions	%
	1	2	3	4	5	6	7	8	9	10	11	12		
A to G	3539	2630	3119	2002	2406	2592	2394	2171	1649	1755	1718	1696	27671	34.5
T to C	3618	2718	3192	1985	2400	2646	2359	2217	1803	1857	1875	1773	28443	35.5
A to T	894	667	769	573	620	715	575	604	501	511	514	503	7446	9.3
C to A	829	614	746	495	515	579	480	491	423	427	421	392	6412	8.0
T to G	776	585	637	525	544	575	497	473	344	383	381	384	6104	7.6
G to C	506	348	439	343	328	345	303	340	253	286	303	257	4051	5.1
Total													**80127**	

Supplementary Table 17. Locations of SSRs relative to genetic marker**s.**

Too large to display. Please see

http://ricelab.plbr.cornell.edu/publications/2005/IRGSP/supplementaltable16.xls

Supplementary Table 18. Information on all SSRs.
Too large to display. Please see
http://ricelab.plbr.cornell.edu/publications/2005/IRGSP/supplementalta

Supplementary Table 19 **Coverage of the pseudomolecules by the draft sequences**

Chromosome	Pseudomolecule (bp)	BGI aligned length[a] (bp)	Coverage (%)	Syngenta aligned length[b] (bp)	Coverage (%)
1	43261740	31010925	71.7	34994029	80.9
2	35954743	25659870	71.4	29339813	81.6
3	36192742	28461838	78.6	28921566	79.9
4	35498469	24328866	68.5	27745686	78.2
5	29737217	22571349	75.9	23550481	79.2
6	30731886	21748471	70.8	24020916	78.2
7	29644043	18646802	62.9	23090121	77.9
8	28434780	20431436	71.9	22569219	79.4
9	22696651	15646883	68.9	17977251	79.2
10	22685906	15204475	67.0	16947181	74.7
11	28386948	17637283	62.1	19395129	68.3
12	27566993	16190624	58.7	21249334	77.1
Total	370792118	257538822	69.5	289800726	78.2
Gene models	37544	22376	59.6	26424	70.4
Supported[c]	9485	6482	68.3	7139	75.3

[a]Aligned length of 50,231contigs of the *O. sativa* ssp. *indica* 93-11 assembly on the IRGSP pseudomolecules requiring matches of at least 80% of the length of the contig and at least 50% identity.

[b]Aligned length of 35,047contigs of the Syngenta assembly of *O. sativa* ssp. *japonica* cv. Nipponbare on the IRGSP pseudomolecules requiring matches of at least 95% coverage and 95% identity.

[c]Gene models supported by coverage of full-length cDNAs for 90% of their length.

Supplementary Table 20 **Comparison of BGI assemblies with IRGSP chromosome 1S**[a]

	BGI 93-11 assembly 93 contigs		BGI Syngenta assembly 70 contigs	
Contigs with homology	71	76.34%	59	84.29%
Duplicate contigs	36	50.70%	6	16.67%
Mis-mapped			11	15.71%
Non-homologous	22	23.66%		
Total contig length (bp)	993,515		848,454	
Average contig length (bp)	10,683		12,121	
Non-redundant coverage (bp)	710,471	82.31%	724,411	81.94%
Overlap (bp)	4,176	0.59%	6,799	0.94%
Mis-matches (bp)	4,108	0.58%	347	0.05%

[a]Comparison with the first 875,786 bp of chromosome 1 pseudomolecule from the telomere of chromosome 1S.

Supplementary Table 21 A **Distribution of CentO sequences in the BGI 93-11 assembly**

Chromosome assignment	Subject start	Subject end	Array length (bp)	CentO copies (No.)
Chr01	1256168	1255350	818	5
Chr01	9043263	9050561	7298	47
Chr01	9186320	9183849	2471	16
Chr01[a]	18569982	18568711	1271	8
Chr01	18624116	18614322	9794	63
Chr01	20350374	20350920	546	4
Chr01	21844437	21846291	1854	12
Chr01	30295598	30297119	1521	10
Chr01	34802712	34806953	4241	27
Chr01	34809809	34808897	912	6
Chr01	38392847	38388111	4736	31
Chr01	40870249	40865995	4254	27
Chr01	40870331	40877675	7344	47
Chr01	40887738	40889385	1647	11
Chr02	4003051	4007367	4316	28
Chr02	14200941	14198813	2128	14
Chr02	14565278	14574523	9245	60
Chr02	14678314	14677370	944	6
Chr02	17796867	17798320	1453	9
Chr02	17809340	17809753	413	3
Chr02	19850232	19854195	3963	26
Chr02	35899130	35901303	2173	14
Chr03	281437	281027	410	3
Chr03	1377354	1382135	4781	31
Chr03	2702221	2696627	5594	36
Chr03	9382323	9384102	1779	11
Chr03	13940122	13941217	1095	7
Chr03	14834636	14830470	4166	27
Chr03	16656112	16660053	3941	25
Chr03	19407814	19409640	1826	12
Chr03	22015728	22017923	2195	14
Chr03	22098541	22093460	5081	33
Chr03	22130775	22130899	124	1
Chr03	22138764	22139081	317	2
Chr03	22153073	22154263	1190	8
Chr03	22157195	22158990	1795	12
Chr03	22167115	22167197	82	1
Chr03	24294585	24297587	3002	19
Chr03	26395800	26393583	2217	14
Chr03	27690907	27683051	7856	51
Chr03	36457961	36461010	3049	20
Chr03	36464095	36464673	578	4
Chr04	1144279	1141426	2853	18
Chr04	7500829	7500708	121	1
Chr04	7506385	7504551	1834	12
Chr04	7519837	7520932	1095	7
Chr04	7541739	7545495	3756	24
Chr04	10169269	10171379	2110	14

Chr04	10261384	10262440	1056	7
Chr04	28468122	28466133	1989	13
Chr04	31512669	31514739	2070	13
Chr06	9860291	9860412	121	1
Chr06	11469301	11465900	3401	22
Chr06	12146383	12148055	1672	11
Chr06	12172750	12173306	556	4
Chr06	14338949	14337345	1604	10
Chr06	15908546	15918312	9766	63
Chr06	16124576	16118167	6409	41
Chr06	16124724	16127603	2879	19
Chr06	19586285	19586747	462	3
Chr06	19588513	19593212	4699	30
Chr07	6549912	6550541	629	4
Chr07	11450228	11452485	2257	15
Chr08	7439696	7443556	3860	25
Chr08	12775077	12773475	1602	10
Chr08	13532245	13534013	1768	11
Chr08	14269758	14271579	1821	12
Chr08	14333346	14330266	3080	20
Chr08	14937002	14936088	914	6
Chr08	14937037	14937284	247	2
Chr08	21487250	21489553	2303	15
Chr08	21822077	21822192	115	1
Chr08	27286179	27286753	574	4
Chr08	27311471	27307908	3563	23
Chr08	27320917	27315883	5034	32
Chr08	27323434	27323846	412	3
Chr08	27324908	27325062	154	1
Chr08	27325348	27326296	948	6
Chr08	28509724	28509851	127	1
Chr09	1315388	1312201	3187	21
Chr09	2104886	2107407	2521	16
Chr09	2613609	2610676	2933	19
Chr09	2976137	2975983	154	1
Chr09	4978842	4978050	792	5
Chr09	15388053	15390390	2337	15
Chr09	15396620	15401683	5063	33
Chr10	4784639	4787205	2566	17
Chr10	4789705	4790963	1258	8
Chr10	6602523	6603245	722	5
Chr10	7464130	7463857	273	2
Chr10	13456190	13448564	7626	49
Chr10	14969094	14970834	1740	11
Chr11	10771594	10773866	2272	15
Chr11	11946953	11948283	1330	9
Chr12	9348020	9344974	3046	20
Chr12	9352184	9350529	1655	11
Chr12	9458337	9458192	145	1
Chr12	9458524	9459748	1224	8
Total			243125	1569
Total non-centromeric			165815	1070
				68.2%

[a]Rows in red are probable centromeric regions.

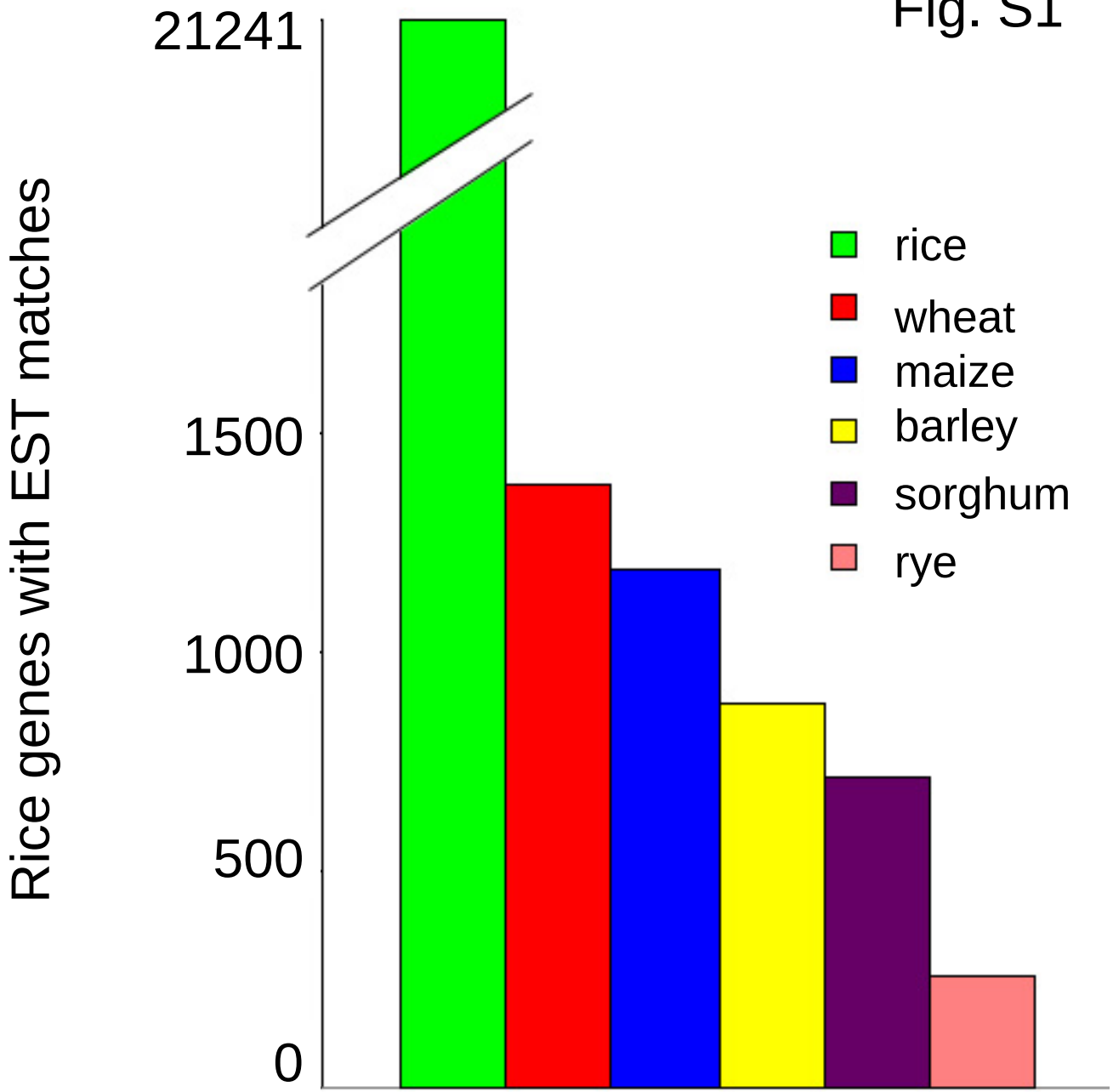
Supplementary Table 21B **Distribution of CentO sequences in the BGI Syngenta assembly**

Chromosome assignment	Subject start	Subject end	Array length (bp)	CentO copies (No.)
Chr01[a]	16731017	16735672	4655	30
Chr01	16738393	16741278	2885	19
Chr01	16745887	16746623	736	5
Chr01	16780879	16790706	9827	63
Chr01	25211958	25210493	1465	9
Chr01	25216170	25213019	3151	20
Chr02	13056387	13055460	927	6
Chr02	13529745	13528220	1525	10
Chr03	252854	252444	410	3
Chr03	3906999	3907126	127	1
Chr03	17720708	17721002	294	2
Chr03	17722657	17728201	5544	36
Chr03	19959184	19961392	2208	14
Chr03	20036563	20036162	401	3
Chr03	20067645	20067770	125	1
Chr03	20099012	20098930	82	1
Chr03	20109917	20108122	1795	12
Chr03	20114786	20113598	1188	8
Chr03	23748094	23749467	1373	9
Chr03	23772663	23776063	3400	22
Chr04	7791498	7794806	3308	21
Chr04	7795363	7796695	1332	9
Chr04	7798649	7804989	6340	41
Chr04	7811886	7814379	2493	16
Chr04	7818720	7823580	4860	31
Chr05	12278501	12282262	3761	24
Chr06	8971387	8971508	121	1
Chr06	14640316	14642546	2230	14
Chr07	10815799	10817384	1585	10
Chr07	10911755	10911887	132	1
Chr07	10911920	10913263	1343	9
Chr07	10918808	10920391	1583	10
Chr07	11458963	11465777	6814	44
Chr07	23154880	23152140	2740	18
Chr08	9437713	9435791	1922	12
Chr08	9443242	9441814	1428	9
Chr08	12176063	12181643	5580	36
Chr08	12273671	12272392	1279	8
Chr08	12799984	12798723	1261	8
Chr08	13500044	13496155	3889	25
Chr08	13500150	13502433	2283	15
Chr08	19353176	19353399	223	1
Chr08	25583844	25583968	124	1
Chr09	1489178	1491035	1857	12
Chr09	2277924	2274987	2937	19
Chr09	2652812	2650745	2067	13
Chr09	2656497	2653667	2830	18
Chr09	2664137	2663077	1060	7

Chr09	2676134	2677518	1384	9
Chr09	2695489	2696408	919	6
Chr09	2726512	2726207	305	2
Chr09	2729195	2729551	356	2
Chr09	2735519	2733547	1972	13
Chr10	6046114	6047077	963	6
Chr10	6049145	6050552	1407	9
Chr10	6247385	6246088	1297	8
Chr10	7254983	7256528	1545	10
Chr10	7979969	7979696	273	2
Chr11	10903315	10908209	4894	32
Chr12	7386026	7387229	1203	8
Chr12	9447045	9453431	6386	41
Chr12	9463727	9466354	2627	17
Chr12	9491249	9496539	5290	34
Chr12	9509827	9514860	5033	32
Chr12	9552810	9550532	2278	15
Chr12	9563131	9560252	2879	19
Chr12	9578031	9573203	4828	31
Chr12	9578140	9581952	3812	25
Total			140321	905
Non-centromeric subtotal			137475	293
				32.4%

[a]Rows in red are probable centromeric regions.

Fig. S1


21241
1500
1000
500
0
Rice genes with EST matches
rice
wheat
maize
barley
sorghum
rye


30
15
0
5
10
15
20
25
30
35
40
45 Mb

Fig. S3a

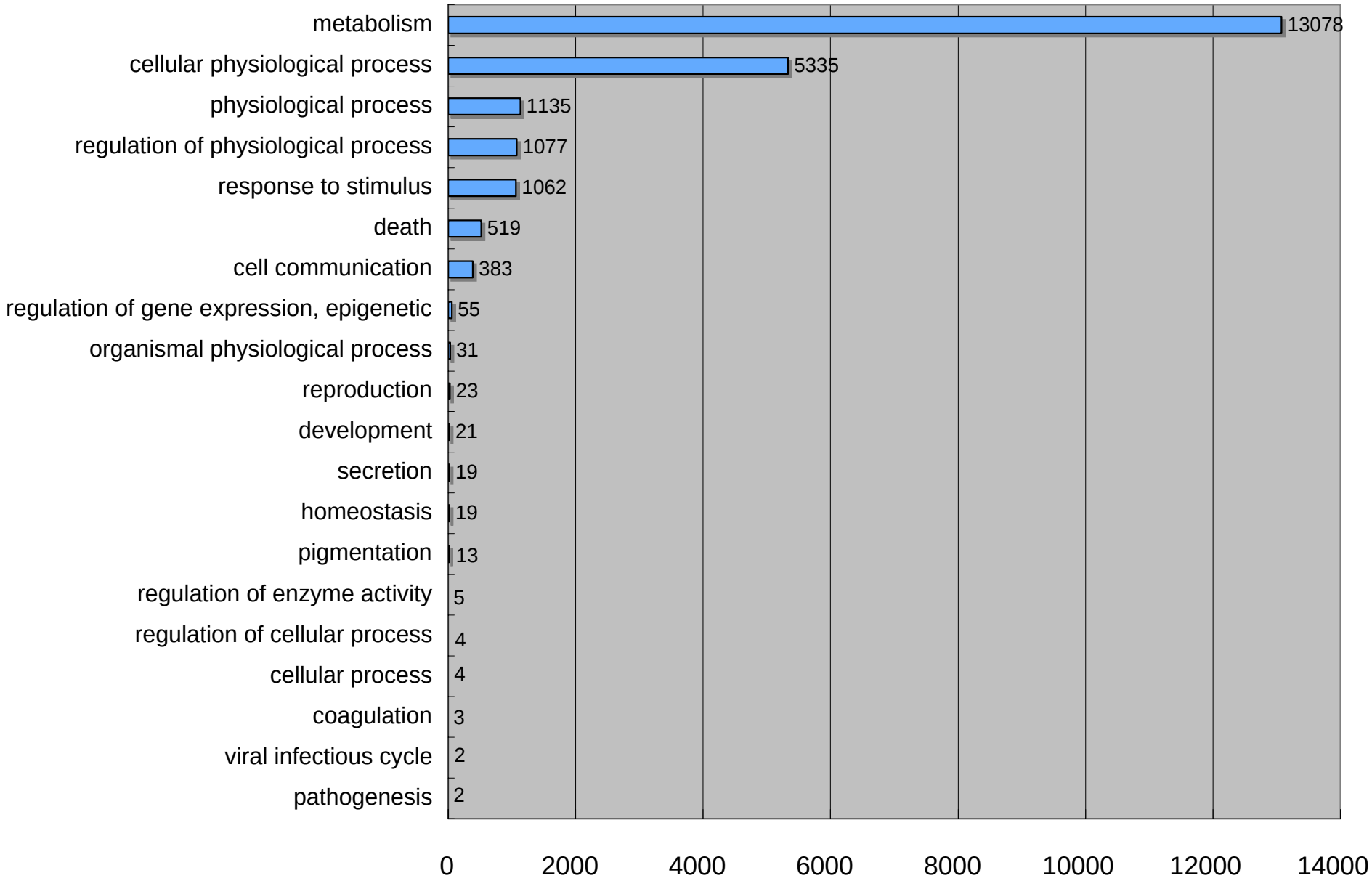
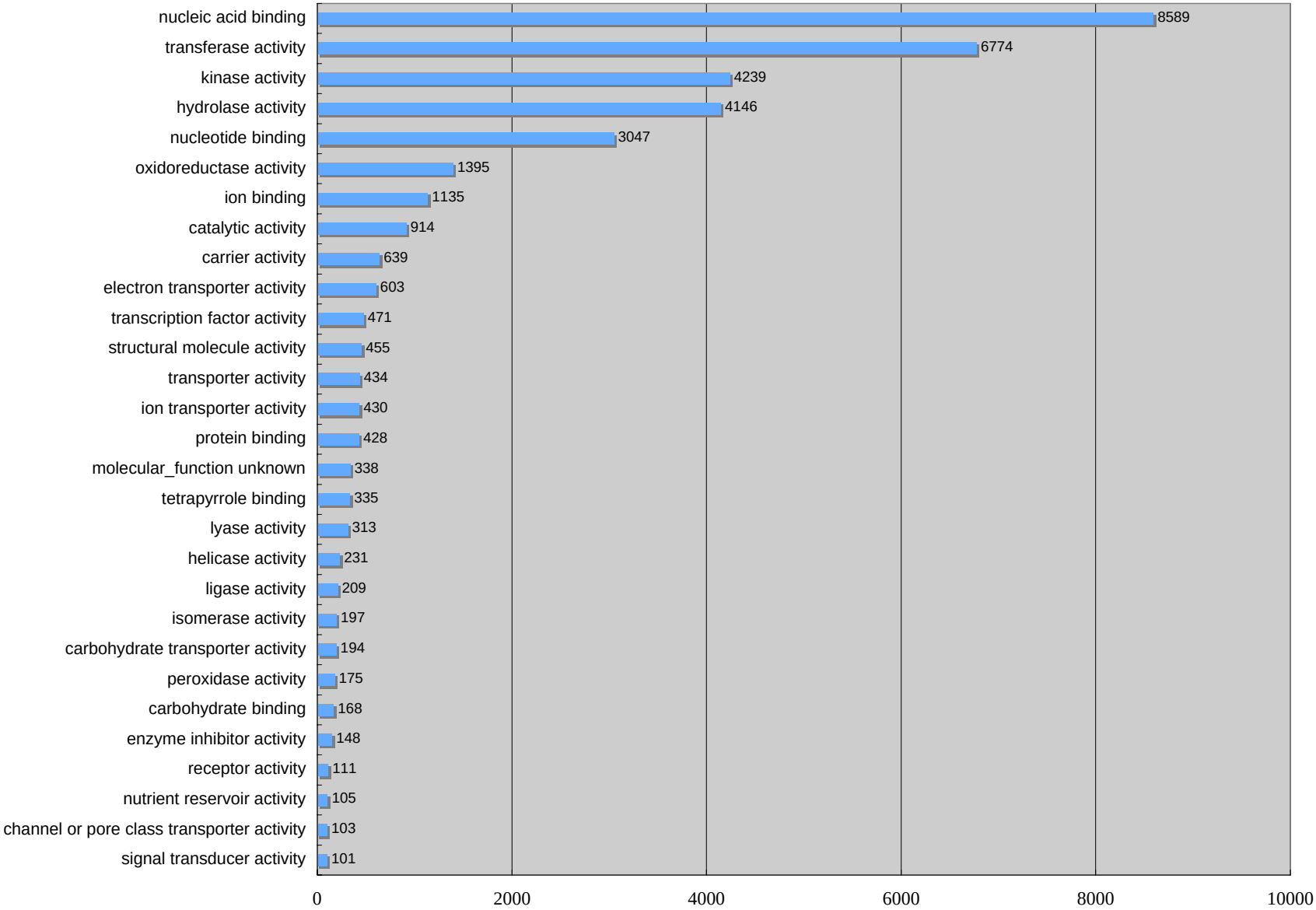
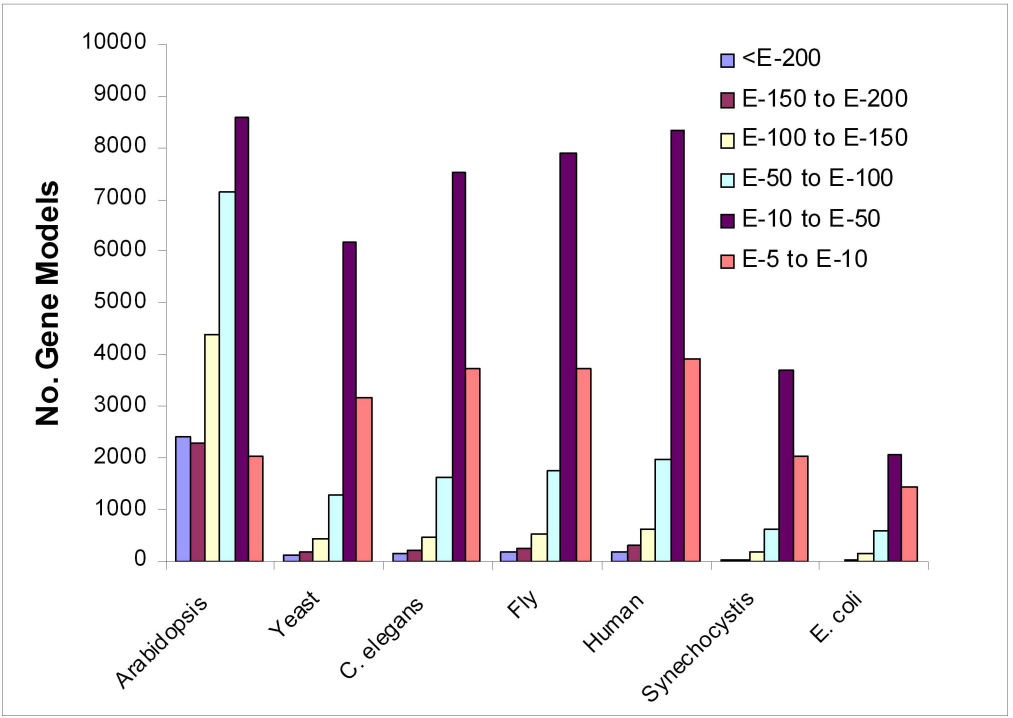
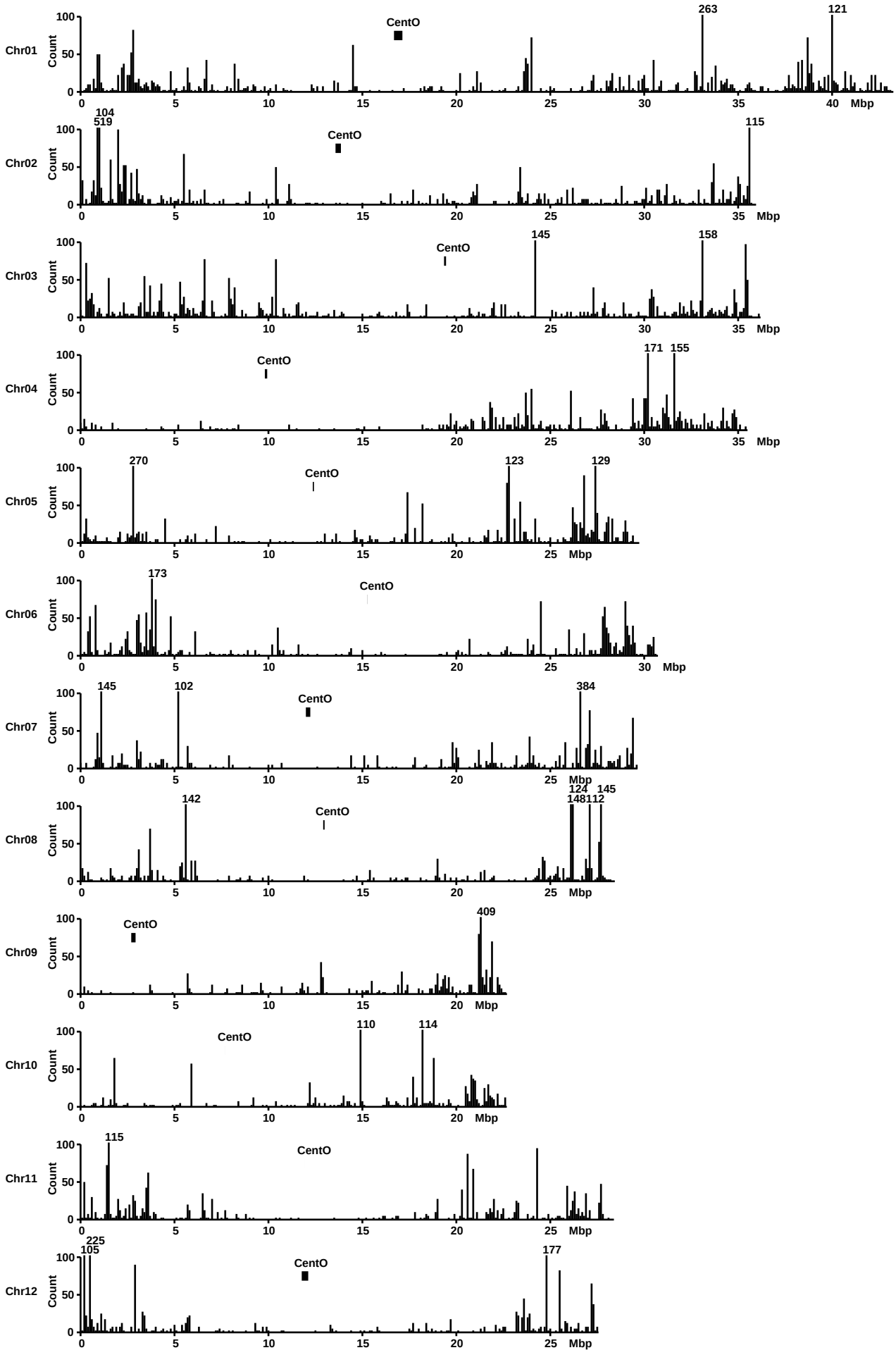
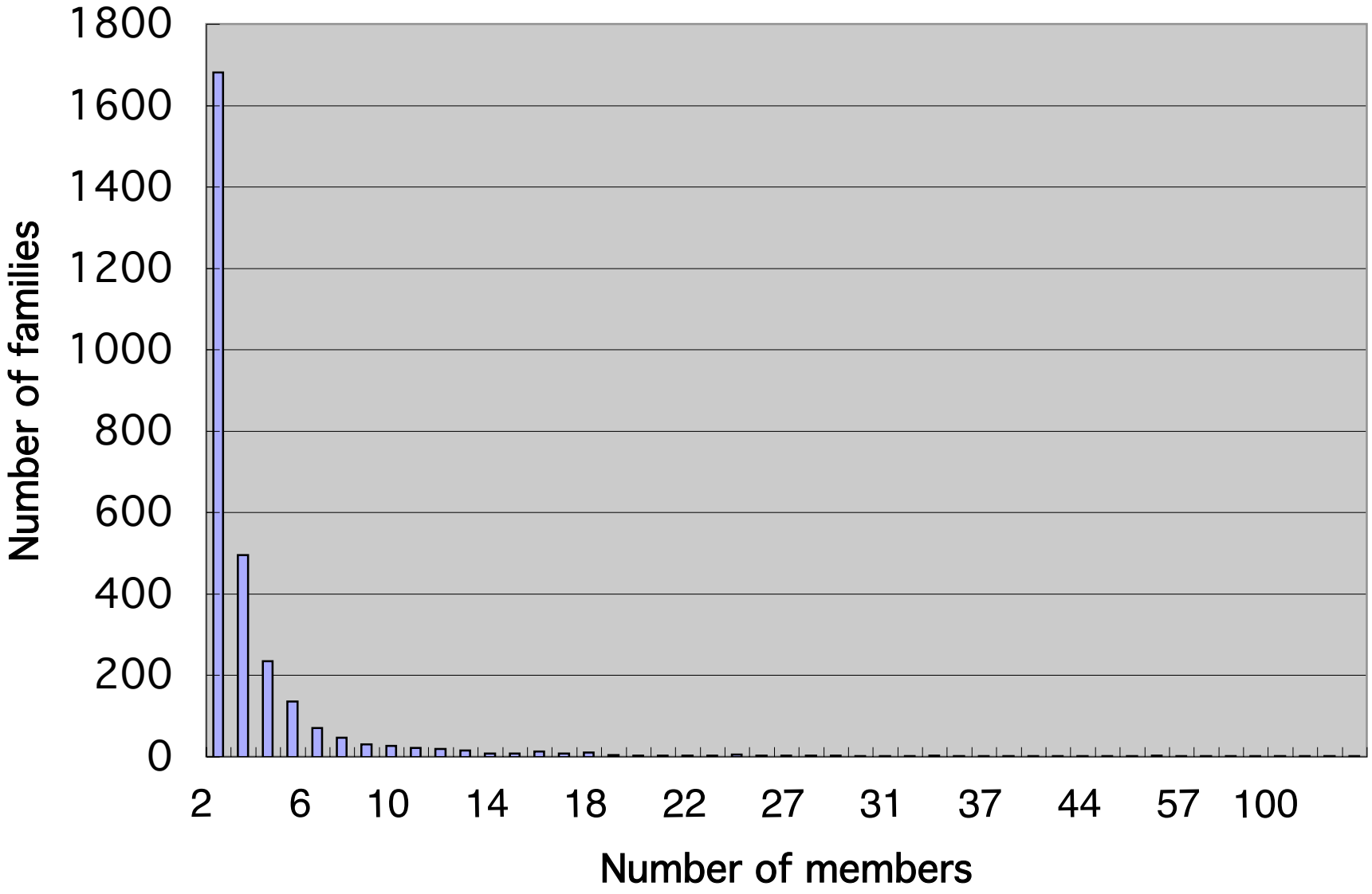

metabolism
13078
cellular physiological process
5335
physiological process
1135
regulation of physiological process
1077
response to stimulus
1062
death
519
cell communication
383
regulation of gene expression, epigenetic
55
organismal physiological process
31
reproduction
23
development
21
secretion
19
homeostasis
19
pigmentation
13
regulation of enzyme activity
5
regulation of cellular process
4
cellular process
4
coagulation
3
viral infectious cycle
2
pathogenesis
2
0
2000
4000
6000
8000
10000
12000
14000

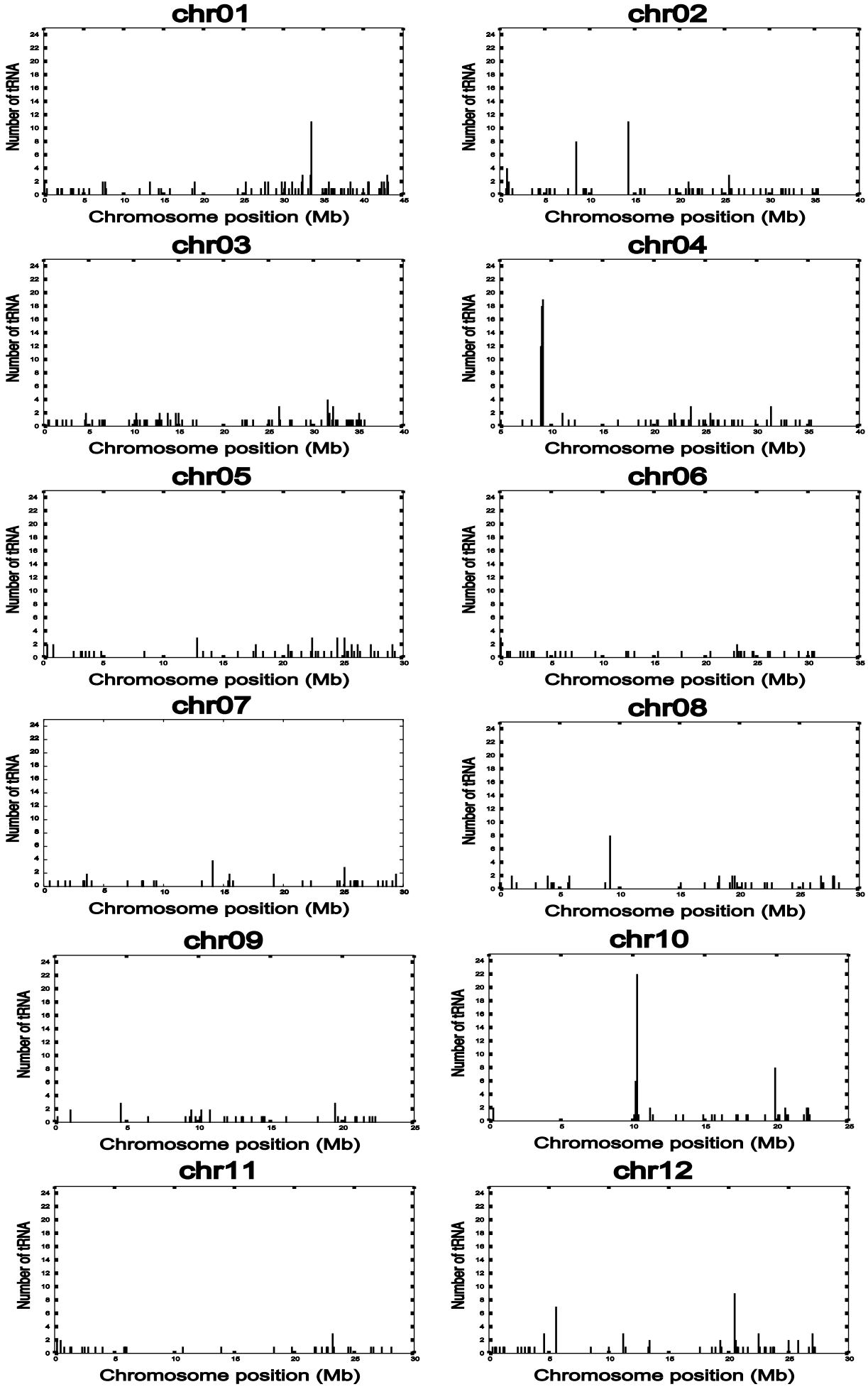
Fig. S3b

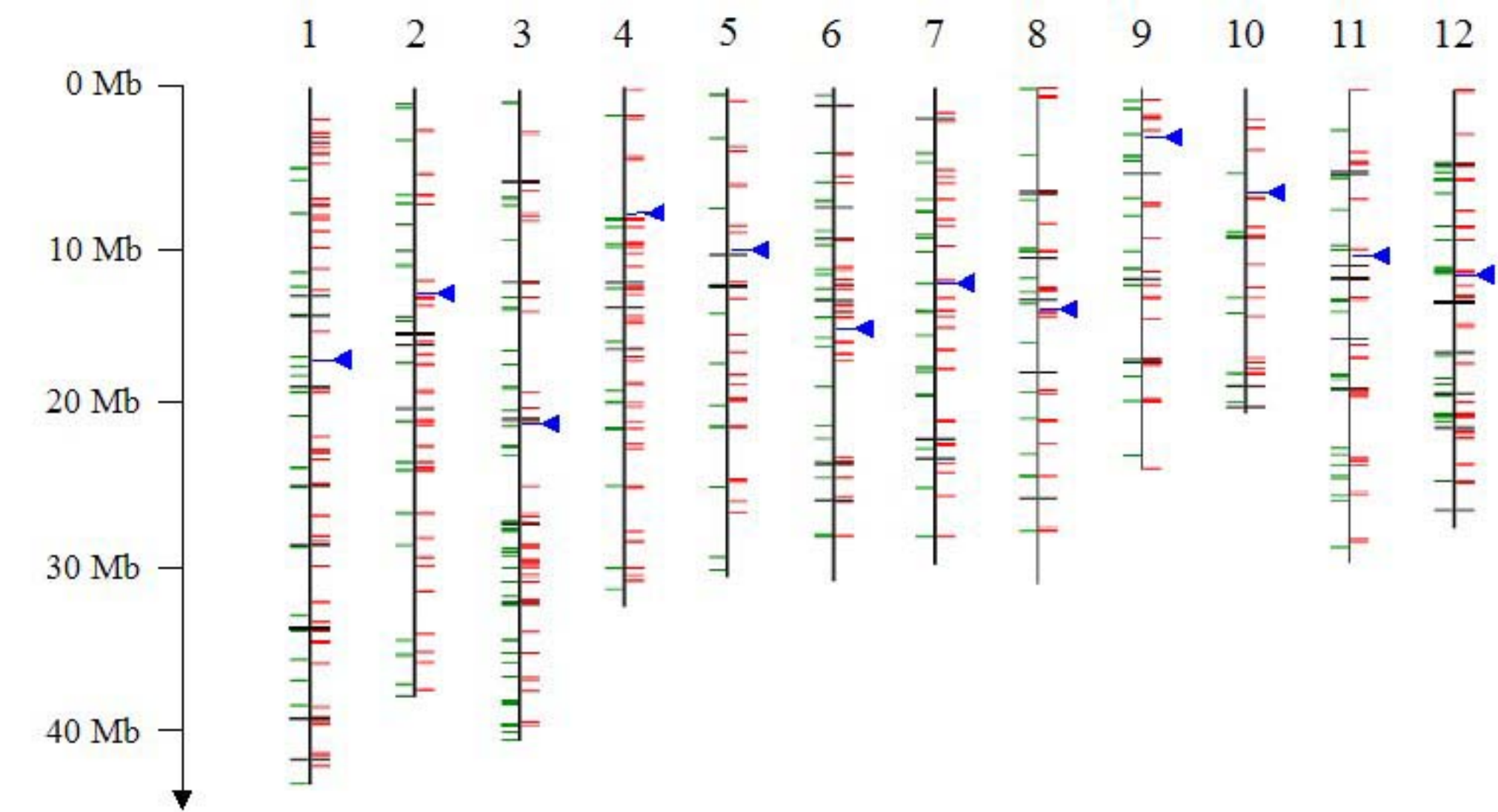

nucleic acid binding 8589
transferase activity 6774
kinase activity 4239
hydrolase activity 4146
nucleotide binding 3047
oxidoreductase activity 1395
ion binding 1135
catalytic activity 914
carrier activity 639
electron transporter activity 603
transcription factor activity 471
structural molecule activity 455
transporter activity 434
ion transporter activity 430
protein binding 428
molecular_function unknown 338
tetrapyrrole binding 335
lyase activity 313
helicase activity 231
ligase activity 209
isomerase activity 197
carbohydrate transporter activity 194
peroxidase activity 175
carbohydrate binding 168
enzyme inhibitor activity 148
receptor activity 111
nutrient reservoir activity 105
channel or pore class transporter activity 103
signal transducer activity 101
0
2000
4000
6000
8000
10000

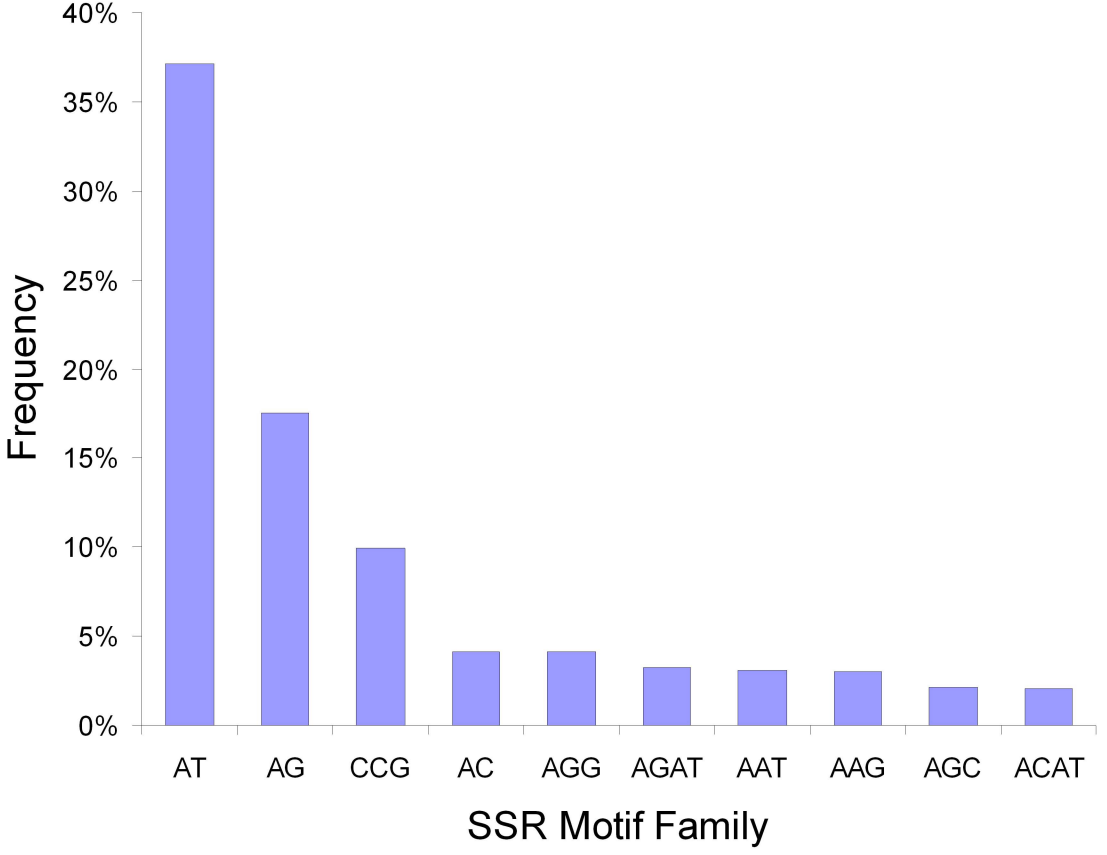

No. Gene Models
10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
<E-200
E-150 to E-200
E-100 to E-150
E-50 to E-100
E-10 to E-50
E-5 to E-10
Arabidopsis
Yeast
C. elegans
Fly
Human
Synechocystis
E. coli

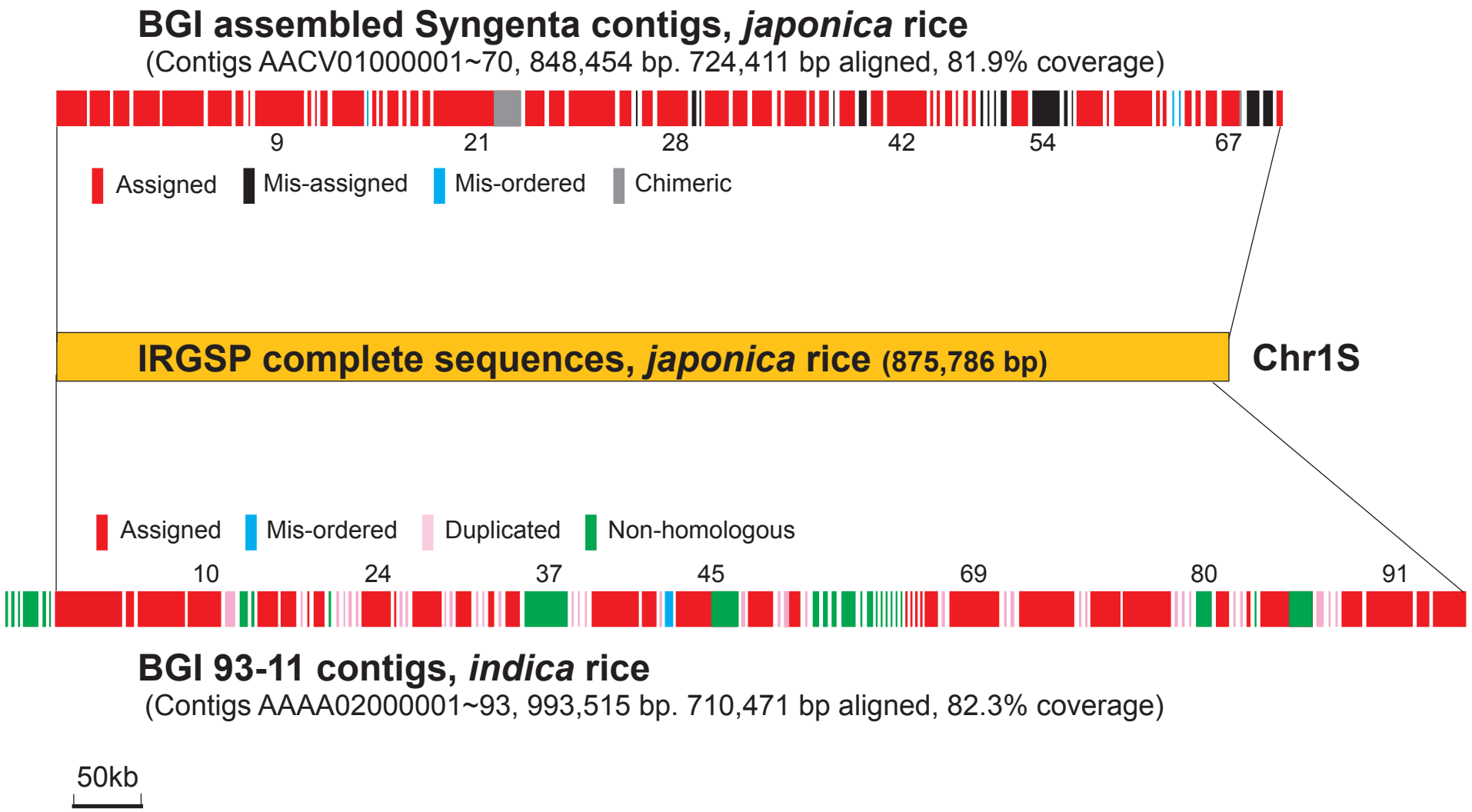

Chr01
Count
CentO
263
121
Mbp
Chr02
104
519
CentO
115
Mbp
Chr03
CentO
145
158
Mbp
Chr04
CentO
171
155
Mbp
Chr05
270
CentO
123
129
Mbp
Chr06
173
CentO
Mbp
Chr07
145
102
CentO
384
Mbp
Chr08
142
CentO
124
145
148
112
Mbp
Chr09
CentO
409
Mbp
Chr10
CentO
110
114
Mbp
Chr11
115
CentO
Mbp
Chr12
225
105
CentO
177
Mbp


1800
1600
1400
1200
1000
800
600
400
200
0
Number of families
2
6
10
14
18
22
27
31
37
44
57
100
Number of members


chr01
chr02
chr03
chr04
chr05
chr06
chr07
chr08
chr09
chr10
chr11
chr12
Number of tRNA
Chromosome position (Mb)


1
2
3
4
5
6
7
8
9
10
11
12
0 Mb
10 Mb
20 Mb
30 Mb
40 Mb


40%
35%
30%
25%
20%
15%
10%
5%
0%
Frequency
AT
AG
CCG
AC
AGG
AGAT
AAT
AAG
AGC
ACAT
SSR Motif Family


BGI assembled Syngenta contigs, japonica rice
(Contigs AACV01000001~70, 848,454 bp. 724,411 bp aligned, 81.9% coverage)
9
21
28
42
54
67
Assigned
Mis-assigned
Mis-ordered
Chimeric
IRGSP complete sequences, japonica rice (875,786 bp)
Chr1S
Assigned
Mis-ordered
Duplicated
Non-homologous
10
24
37
45
69
80
91
BGI 93-11 contigs, indica rice
(Contigs AAAA02000001~93, 993,515 bp. 710,471 bp aligned, 82.3% coverage)
50kb