viruses were injected to follicles on both wings for later studies. Chickens were raised in cages and observed on a daily basis over a two-month period. The regenerated feathers were plucked and examined with a dissection or scanning electron micrograph microscope for abnormalities compared with normal primary remiges.

Histology and *in situ* hybridization

Paraffin sections (5 μm) were stained with haematoxylin and eosin or prepared for *in situ* hybridization following routine procedures[26]. Cryostat sections (10 μm) were stained with X-gal. TUNEL staining was performed using a kit (Roche). Nonradioactive wholemount or section *in situ* hybridization or section *in situ* hybridization was performed according to the protocol described[22,26]. After hybridization, sections were incubated with an anti-digoxigenin Fab conjugated to alkaline phosphatase (Boehringer Mannheim). Colour was detected by incubating with a Boehringer Mannheim purple substrate (Roche).

Received 5 June; accepted 10 October 2002; doi:10.1038/nature01196.
Published online 30 October 2002.

1. Lucas, A. M. & Stettenheim, P. R. (eds) *Avian Anatomy – Integument. Agricultural Handbook 362: Agricultural Research Services* (US Department of Agriculture, Washington DC, 1972).
2. Chuong, C.-M. The making of a feather: Homeoproteins, retinoids and adhesion molecules. *BioEssays* **15,** 513–521 (1993).
3. Feduccia, A. *The Origin and Evolution of Birds* 2nd edn (Yale Univ. Press, New Haven, Connecticut, 1999).
4. Chatterjee, S. *The Rise of Birds* (John Hopkins Univ. Press, Baltimore, Maryland, 1997).
5. Regal, P. J. The evolutionary origin of feathers. *Q. Rev. Biol.* **50,** 35–66 (1975).
6. Chen, P. J., Dong, Z. M. & Shen, S. N. An exceptionally well-preserved theropod dinosaur from the Yixian Formation of China. *Nature* **391,** 147–152 (1998).
7. Xu, X., Tang, Z. L. & Wang, X. L. A therizinorsauroid dinosaur with integumentary structures from China. *Nature* **399,** 350–354 (1999).
8. Jones, T. D. *et al.* Nonavian feathers in a late Triassic archosaur. *Science* **288,** 2202–2205 (2000).
9. Prum, R. O. Longisquama fossil and feather morphology. *Science* **291,** 1899–1902 (2001).
10. Zhang, F. & Zhou, Z. A primitive enantiornithine bird and the origin of feathers. *Science* **290,** 1955–1959 (2000).
11. Xu, X., Zhou, Z. & Prum, R. O. Branched integumental structures in Sinornithosaurus and the origin of feathers. *Nature* **410,** 200–204 (2001).
12. Ji, Q., Currie, P. J., Norell, M. A. & Ji, S. A. Two feathered dinosaurs from northeast China. *Nature* **393,** 753–761 (1998).
13. Ji, Q., Norell, M. A., Gao, K. Q., Ji, S. A. & Ren, D. The distribution of integumentary structures in a feathered dinosaur. *Nature* **410,** 1084–1088 (2001).
14. Norell, M. *et al.* Modern feathers on a non-avian dinosaur. *Nature* **416,** 36–37 (2002).
15. Morgan, B. A. & Fekete, D. M. Manipulating gene expression with replication-competent retroviruses. *Methods Cell Biol.* **51,** 185–218 (1996).
16. Prum, R. O. Development and evolutionary origin of feathers. *J. Exp. Zool.* **285,** 291–306 (1999).
17. Prum, R. O. & Williamson, S. Theory of the growth and evolution of feather shape. *J. Exp. Zool.* **291,** 30–57 (2001).
18. Chuong, C.-M., Chodankar, R., Widelitz, R. B. & Jiang, T.-X. Evo-devo of feathers and scales: building complex epithelial appendages. *Curr. Opin. Genet. Dev.* **10,** 449–456 (2000).
19. Chuong, C.-M. *et al.* Dinosaur's feather and Chicken's tooth? Tissue engineering of the integument. John Ebbling lecture. *Eur. J. Dermatology* **11,** 286–292 (2001).
20. Chuong, C.-M. (ed.) *Molecular Basis of Epithelial Appendage Morphogenesis* (Landes Bioscience, Austin, 1998).
21. Hogan, B. L. M. Morphogenesis. *Cell* **96,** 225–233 (1999).
22. Jung, H.-S. *et al.* Local inhibitory action of BMPs and their relationships with activators in feather formation: implications for periodic patterning. *Dev. Biol.* **196,** 11–23 (1998).
23. Dudley, A. T. & Tabin, C. J. Constructive antagonism in limb development. *Curr. Opin. Genet. Dev.* **10,** 387–392 (2000).
24. Jiang, T.-X., Jung, H.-S., Widelitz, R. B. & Chuong, C.-M. Self organization of periodic patters by dissociated feather mesenchymal cells and the regulation of size, number and spacing of primordia. *Development* **126,** 4997–5009 (1999).
25. Harris, M. P., Fallon, J. F. & Prum, R. O. Shh-Bmp2 signaling module and the evolutionary origin and diversification of feathers. *J. Exp. Zool.* **294,** 160–176 (2002).
26. Ting-Berreth, S. A. & Chuong, C.-M. Sonic hedgehog in feather morphogenesis: induction of mesenchymal condensation and association with cell death. *Dev. Dyn.* **207,** 157–170 (1996).
27. Cooper, M. K., Porter, J. A., Young, K. E. & Beachy, P. A. Teratogen-mediated inhibition of target tissue response to Shh signaling. *Science* **280,** 1603–1607 (1998).
28. Calabretta, R., Nolfi, S., Parisi, D. & Wagner, G. P. Duplication of modules facilitates the evolution of functional specialization. *Artificial Life* **6,** 69–84 (2000).
29. Chuong, C.-M. & Edelman, G. M. Expression of cell adhesion molecules in embryonic induction. II. Morphogenesis of adult feathers. *J. Cell Biol.* **101,** 1027–1043 (1985).
30. Gill, F. B. *Ornithology,* 2nd edn (Freeman, New York, 1994).

Supplementary Information accompanies the paper on *Nature's* website (http://www.nature.com/nature).

Acknowledgements We thank M. Ramos for help in preparing the manuscript; and R. Prum for critical comments on the manuscript. Figure 1b is modified from ref. 1. This work is supported by grants from the National Institute of Arthritis and Musculoskeletal and Skin Diseases, USA, and the National Science Foundation to C.-M.C., and a National Cancer Institute grant to R.B.W.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to C.-M.C. (e-mail: chuong@pathfinder.usc.edu).

The genome sequence and structure of rice chromosome 1

Takuji Sasaki*, Takashi Matsumoto*, Kimiko Yamamoto*, Katsumi Sakata*, Tomoya Baba*, Yuichi Katayose*, Jianzhong Wu*, Yoshihito Niimura†, Zhukuan Cheng‡, Yoshiaki Nagamura*, Baltazar A. Antonio*, Hiroyuki Kanamori*, Satomi Hosokawa*, Masatoshi Masukawa*, Koji Arikawa*, Yoshino Chiden*, Mika Hayashi*, Masako Okamoto*, Tsuyu Ando*, Hiroyoshi Aoki*, Kohei Arita*, Masao Hamada*, Chizuko Harada*, Saori Hijishita*, Mikiko Honda*, Yoko Ichikawa*, Atsuko Idonuma*, Masumi Iijima*, Michiko Ikeda*, Maiko Ikeno*, Sachie Ito*, Tomoko Ito*, Yuichi Ito*, Yukiyo Ito*, Aki Iwabuchi*, Kozue Kamiya*, Wataru Karasawa*, Satoshi Katagiri*, Ari Kikuta*, Noriko Kobayashi*, Izumi Kono*, Kayo Machita*, Tomoko Maehara*, Hiroshi Mizuno*, Tatsumi Mizubayashi*, Yoshiyuki Mukai*, Hideki Nagasaki*, Marina Nakashima*, Yuko Nakama*, Yumi Nakamichi*, Mari Nakamura*, Nobukazu Namiki*, Manami Negishi*, Isamu Ohta*, Nozomi Ono*, Shoko Saji*, Kumiko Sakai*, Michie Shibata*, Takanori Shimokawa*, Ayahiko Shomura*, Jianyu Song*, Yuka Takazaki*, Kimihiro Terasawa*, Kumiko Tsuji*, Kazunori Waki*, Harumi Yamagata*, Hiroko Yamane*, Shoji Yoshiki*, Rie Yoshihara*, Kazuko Yukawa*, Huisun Zhong*, Hisakazu Iwama†, Toshinori Endo§, Hidetaka Ito§, Jang Ho Hahn‖ Ho-Il Kim‖, Moo-Young Eun‖, Masahiro Yano*, Jiming Jiang‡ & Takashi Gojobori†

** Rice Genome Research Program, National Institute of Agrobiological Sciences, and Institute of the Society for Techno-innovation of Agriculture, Forestry and Fisheries, 1-2, Kannondai 2-chome, Tsukuba, Ibaraki 305-8602, Japan*
† Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, Mishima 411-8540, Japan
‡ Department of Horticulture, University of Wisconsin-Madison, Madison, Wisconsin 53706, USA
§ Department of Bioinformatics, Tokyo Medical and Dental University, 1-5-45 Yushima, Bunkyo-ku, Tokyo 113-8510, Japan
‖ Rice Genome Sequencing Project, National Institute of Agricultural Science and Technology, RDA, 249 Seodun-dong, Suwon 441-707, Korea

The rice species *Oryza sativa* is considered to be a model plant because of its small genome size, extensive genetic map, relative ease of transformation and synteny with other cereal crops[1–4]. Here we report the essentially complete sequence of chromosome 1, the longest chromosome in the rice genome. We summarize characteristics of the chromosome structure and the biological insight gained from the sequence. The analysis of 43.3 megabases (Mb) of non-overlapping sequence reveals 6,756 protein coding genes, of which 3,161 show homology to proteins of *Arabidopsis thaliana*, another model plant. About 30% (2,073) of the genes have been functionally categorized. Rice chromosome 1 is (G + C)-rich, especially in its coding regions, and is characterized by several gene families that are dispersed or arranged in tandem repeats. Comparison with a draft sequence[5] indicates the importance of a high-quality finished sequence.

Rice has been studied extensively by molecular genetics and constitutes one of the best characterized crop plants with a fine genetic map of 3,267 markers (http://rgp.dna.affrc.go.jp/publicdata/geneticmap2000/index.html)[1], a yeast artificial chromosome (YAC) physical map with 80.8% coverage[2], sequences for about 10,000 unique expressed sequence tags (ESTs)[3], and a transcriptional map indicating the placement of 6,591 unique ESTs[2]. The Rice Genome Research Program (RGP) in Japan launched its rice genome sequencing project in 1998. It is a partner of the International Rice Genome Sequencing Project (IRGSP), which involves ten countries in Asia, North America, South America and Europe that are working towards the immediate release of high-quality sequence data to the public domain[4]. The draft sequences of the two

main subspecies of rice, *japonica* and *indica*, have been reported[5,6]. Both studies were based on whole-genome shotgun sequencing rather than on the clone-by-clone approach of the IRGSP. Although the release of the draft sequence is of immense scientific value, many challenges in rice genomics demand the availability of a complete, accurate, map-based rice genome sequence.

We determined the sequence of chromosome 1 from 390 overlapping phage (P1)-derived artificial chromosome (PAC) and bacterial artificial chromosome (BAC) clones and assembled it into nine contigs (Fig. 1). The longest contig is 14.4 Mb and spans positions 106.2 centimorgans (cM) to 157.1 cM on the molecular genetic map. Among the eight remaining gaps, gap 4, located at 73.4 cM, corresponds to a portion of the centromeric region and is estimated to be about 1,400 kilobases (kb) by the pachytene fluorescence *in situ* hybridization (FISH) method[7]. PAC/BAC clones adjacent to this gap contain copies of the rice centromere-specific sequence RCS2 (ref. 8). Two PAC clones, P0402A09 and P0020E09, are localized to the most distal ends of the short arm and the long arm, and their map positions have been verified by pachytene FISH using pAtT4 (ref. 9), a telomeric clone of *Arabidopsis* (Supplementary Fig. 1). This indicates that our physical map extends to within less than 50 kb of the telomeres. Integration of the PAC/BAC physical mapping with the results from fibre FISH gives a total length of 45.7 Mb for chromosome 1, corresponding to 181.8 cM on the genetic map, excluding the telomeres.

Statistics for the nucleotide sequence of rice chromosome 1 are summarized in Table 1. The non-overlapping sequence covers 43,276,883 nucleotides. In this sequence, 6,756 genes were either identified or predicted. Thus, the average gene density of chromosome 1 is about one gene per 6.4 kb. If this distribution is assumed to be similar throughout the whole genome, then the total number of genes in the rice genome (400 Mb) is roughly 62,500. This number is 2.5 times larger than the gene total of *Arabidopsis*[10]. But this difference might easily be the result of an overestimate of rice genes, because it assumes that there is a uniform distribution of genes along the chromosomes.

Cytogenetic analysis has indicated clear differences in the content of heterochromatin in each of the 12 rice chromosomes, and chromosome 1 shows the least amount of heterochromatic material[11]. The average exon size is comparable to that of *Arabidopsis*, but the average intron size is about 3.6 times larger. This means that, although the longer introns engender larger gene sizes in rice, the average transcriptome size is similar in both species. The G + C content of coding and noncoding regions in rice is higher than in *Arabidopsis*—the rice coding regions are especially (G + C)-rich. This characteristic is reflected by the biased usage of G/C at the third position of codons within predicted genes (Supplementary Table 1). Buoyant density experiments have shown that rice genes are localized in (G + C)-rich islands that occupy 24% of the genome[12]. When we plotted the average G + C values against chromosomal position in chromosome 1, however, we did not detect any CpG islands, indicating a neutral nucleotide distribution. The ratios of physical to genetic distance on the short and the long arms are 214 kb cM^{-1} ($r^2 = 0.983$) and 288 kb cM^{-1} ($r^2 = 0.976$), respectively, suggesting that the rate of recombination differs along the two arms of the chromosome.

We compared our finished sequence (493,729 bp from the distal

Table 1 **Compositional analysis of the sequence of rice chromosome 1**

Overall physical length	45,745,883 bp
Short arm	17,081,313 bp
Long arm	27,264,570 bp
Total length of eight gaps	2,469,000 bp
Non-overlapping sequence	43,276,883 bp
Base composition (% GC)	
Overall	43.8%
Coding region	58.2%
Noncoding region	40.7%
Predicted gene number	6,756
Gene density	6.4 kb per gene
Average gene size	3.4 kb
Exons	
Size of exons	1.1 kb per gene
Number of exons per gene	4.8
Introns	
Size of introns	2.3 kb per gene
Number of introns per gene	3.8
Repetitive sequences	
Number of retrotransposons (class I)	3,235
Number of DNA transposons (class II)	
Autonomous type	6,985
Nonautonomous type (MITEs)	14,106
Total number of repeats	24,326

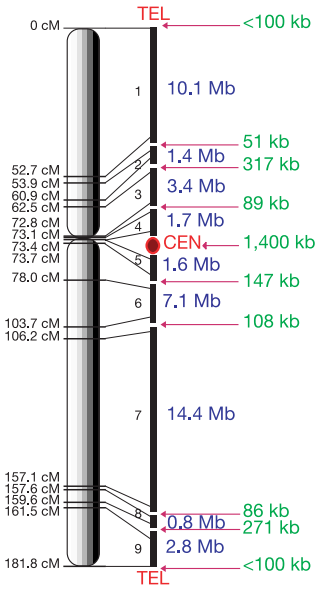

Figure 1 Physical map of rice chromosome 1. Positions of the PAC/BAC contigs are indicated by black bars. Purple numbers indicate the physical distances that were calculated on the basis of the nucleotide sequence length of each contig. A representation of the genetic map of chromosome 1 is shown on the left with the positions of the genetic markers found nearest the end of each contig. The centromeric region is shown as a red circle. The green numbers show the gap sizes as measured by fibre FISH and pachytene FISH.

end of the short chromosome arm) with 127,550 *indica* sequence contigs assembled from the whole-genome shotgun sequences of the Beijing Genomics Institute (BGI, http://btn.genomics.org.cn/rice/) using the *japonica* sequence as a query for basic BLASTN (basic local alignment search tool) analysis (Fig. 2). We could detect the corresponding *indica* sequence in about 78% of the whole region. But there were 65 gaps in the aligned contigs, and a total of 110,389 bases (22%) of *japonica* sequence could not be identified in the *indica* assembly. This may partly reflect the sequence difference between the two subspecies, although some artefacts in the whole-genome shotgun assembly cannot be ruled out. Among the 96 predicted genes in this region of the completed *japonica* sequence, 55 genes are intact, 33 genes are partially predicted and 8 genes are not predicted in the corresponding *indica* draft sequence. Relative identities near the repeat (retrotransposon-like) regions are lower than in the other regions, indicating a misassembly in the sequence.

Direct comparison with the *japonica* draft sequence could not be made because the sequence data are not in the public domain. But previously, 4,467 genes were predicted from a set of 99 BAC contigs assigned to chromosome 1 (ref. 6). It is likely that an estimated 2,835–4,211 gaps (either 63 gaps per megabase or 10% of 42,109 total gaps) for this chromosome prevented an accurate prediction of the number of genes. Not surprisingly, only half of the genes predicted contain complete coding regions. In addition, no basis was provided for the assignment of genes to chromosome[6].

We used an automated annotation system, RiceGAAS[13], to characterize the gene composition of chromosome 1 (Supplementary Fig. 2, http://RiceGAAS.dna.affrc.go.jp/chromosome1/). The distribution of genes along both arms of the chromosome indicates higher density (18–19 genes per 100 kb) in distal as compared with proximal regions (10–12 genes per 100 kb). This was verified by experimental results obtained by mapping 977 expressed sequence tags on to chromosome 1 (ref. 2). Among the 6,756 predicted genes, 2,073 (31%) were functionally characterized by homology to known proteins using BLASTP, whereas 69% of the predicted genes corresponded to proteins with no known function (Table 2). The protein signature search program InterPro detected protein domains in 3,660 (54%) of the total predicted genes (see http://RiceGAAS.dna.affrc.go.jp/chromosome1/). In particular, 1,170 (33%) of 3,600 hypothetical proteins showed domain homology, suggesting that these proteins may correspond to newly identified proteins in rice. BLASTN analysis was done using the cereal EST entries from the EST database at the National Institute for Biotechnology Information (NCBI). Exon regions from all predicted genes were used as queries, and 546,723 unclustered ESTs from wheat, maize, barley and sorghum were searched using a threshold probability value of 10^{-5}. A total of 2,985 predicted genes, including 756 hypothetical proteins, have cereal homologues. Thus, among the 6,756 predicted genes, 4,803 (71%) show some evidence of homology to a domain, a functional site, a cereal EST or a protein.

Table 2 **Functional classification of the proteins encoded on rice chromosome 1**

Total proteins*	6,756
Class I	40
Class II	799
Class III	1,234
Class IV	870
Class V	213
Class VI	3,600
Functional classification	
Cellular metabolism	421 (6.23%)
Signal transduction	365 (5.40%)
Transcription	255 (3.77%)
Cell rescue, defence	215 (3.18%)
Transport facilitation	172 (2.55%)
Cellular organization	132 (1.95%)
Protein destination	108 (1.60%)
Protein synthesis	85 (1.26%)
Energy	83 (1.23%)
Cell growth	62 (0.92%)
Development	35 (0.52%)
Cellular transport	28 (0.41%)
Cellular biogenesis	16 (0.24%)
Classification not yet clear-cut	13 (0.19%)
Ionic homeostasis	9 (0.13%)
Organism-specific	3 (0.04%)
Unclassified	4,754 (70.37%)

*Class I, complete match to a rice protein; class II, strong similarity to a known protein, from the N to the C terminus; class III, similarity to less than 50% of the full length of the target protein; class IV, similarity to a protein of unknown function predicted by genome sequence annotation; class V, match to an EST nucleotide sequence but not to a known protein; class VI, no hit at a threshold probability score of 10^{-20} to any entry in the database.

The predicted proteins found on chromosome 1 were categorized into gene families by BLASTP, using a threshold probability score of 10^{-20} over more than 50% of the length of the gene. The most abundant gene family was the serine/threonine receptor kinase family with 132 members distributed along the chromosome (Fig. 3a). A cluster of this gene family was observed at the distal end of the short arm, although some members of the cluster seemed to be pseudogenes. The highest number of tandem repeats detected at a single site was a cluster of ten copies of the hypothetical gene family located on the short arm of chromosome 1. These results are summarized in Fig. 3b, which shows a dot matrix plot of chromosome 1, indicating the predicted genes with significant homology to a given gene. On this plot, which disregards self-homology, a clear diagonal line was obtained, indicating that a significant number of

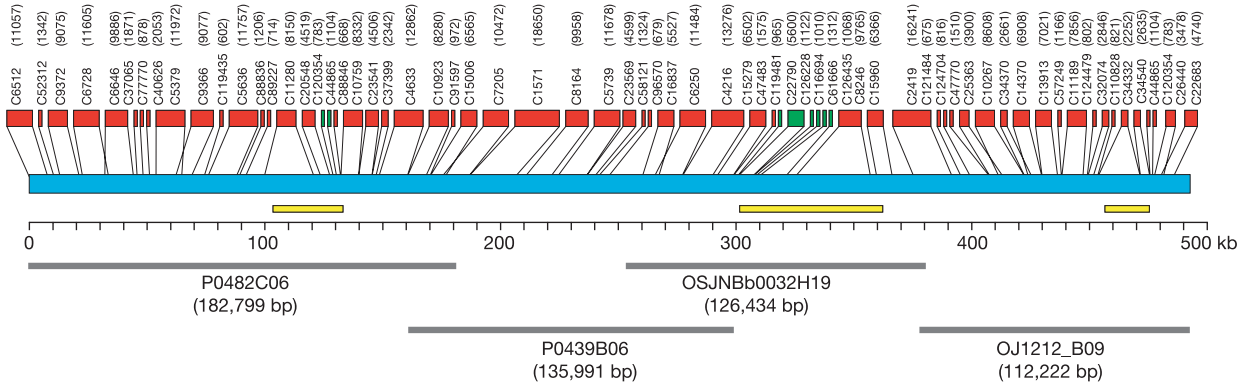

Figure 2 Comparison between the Nipponbare finished sequence and the *indica* draft sequence. Our finished continuous sequence (493,729 bp from near the distal end of the short arm) was used as query. All of the *indica* 93-11 sequence contigs assembled from the whole-genome shotgun sequences from the BGI were downloaded from their website and searched by BLASTN. The highest rank of the hit contigs was aligned to our Nipponbare sequence. Coloured bars represent the following: blue, RGP sequence data; grey, PAC/BAC clones; red, BGI sequence data with >95% identity; green, BGI sequence data with 90–95% identity; and yellow, regions containing repetitive sequences. Numbers in parentheses correspond to the length of PAC/BAC clones and to BGI sequence contigs in bp.

genes are duplicated and arrayed in tandem.

To determine whether any of the proteins on rice chromosome 1 are not present in *Arabidopsis*, the 6,756 predicted proteins were queried in BLASTP searches against all the *Arabidopsis* proteins in the Munich Information Center for Protein Sequence (MIPS) database using a threshold probability score of 10^{-5}. Among 3,161 positive queries, 824 showed strong similarities (probability value less than 10^{-100}) to proteins found in *Arabidopsis*, whereas 3,595 sequences (53%) did not have positive BLASTP hits with predicted *Arabidopsis* proteins at a probability threshold of 10^{-5}. Only 27 of these sequences had homology to known proteins and among them, only Bowman–Birk trypsin inhibitor and cytochrome *f* (chloroplast) were clearly found in rice chromosome 1. This suggests that almost all of the known proteins found in rice chromosome 1 are also found in *Arabidopsis*. Among the hypothetical proteins, 3,051 genes have no counterpart in *Arabidopsis* and 442 (15%) genes have grass orthologues. Analysis of the draft sequence also showed that half of the predicted genes have no homologues in Arabidopsis[5,6]. Although many of these hypothetical genes could be artefacts resulting from prediction errors, functional characterization of these genes in the future may identify grass-specific or even rice-specific genes.

We also observed rice chloroplast genes in sequential order on the chromosomal DNA. For example, at 149.1 cM we identified 3,564 bp of sequence that matched the rice chloroplast sequence with only a 3-bp difference. This sequence contains three genes[14], PSII cytochrome b_{559}, cytochrome *f* and the chloroplast envelope membrane protein ORF230. We also detected 85 putative transfer RNA genes using tRNAscan SE[15]. Analysis of the retrotransposable elements and DNA intermediate transposons, including miniature inverted-repeats transposable elements (MITEs)[16], using RepeatMasker is given in Table 1 and summarized in Supplementary Fig. 3. MITEs have a tendency to be dispersed along the chromosome, whereas the retrotransposons and other autonomous type DNA-mediated transposable elements are clustered in the pericentromeric region. Among retroelements, Ty3/*Gypsy*-type elements are the most frequent (2,157), followed by Ty1/*Copia*-type elements (384). The sum of the lengths of these three repetitive elements is 6.0 Mb, corresponding to 13% of chromosome 1.

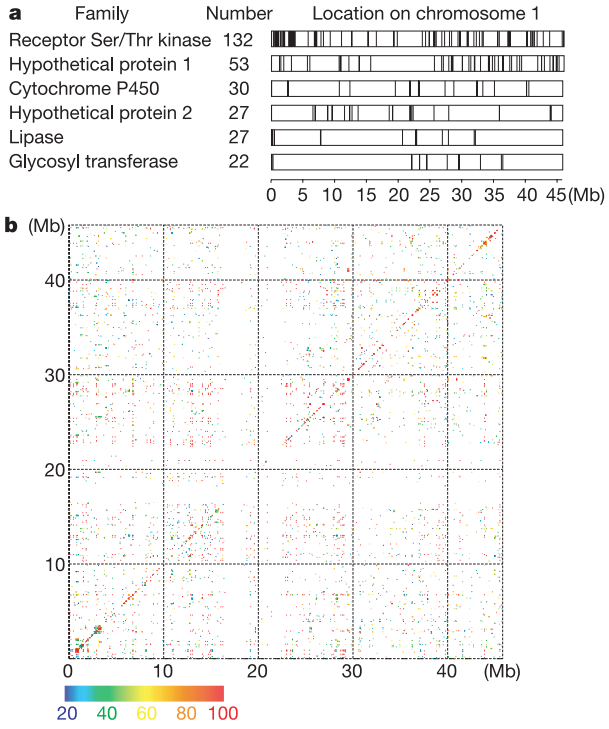

Figure 3 Analysis of gene families and gene clusters. **a**, Distribution of gene families on chromosome 1. For each category, BLASTP was carried out against all of the chromosome 1 gene products. Proteins that had a threshold probability (*E*) value of less than 10^{-20} and that also showed strong homology for more than half of their total length were grouped as a gene family. The six largest gene families are shown. Vertical lines on each bar indicate the positions of gene family members. The bars are oriented with the distal end of the short arm of the chromosome to the left. The scale at the bottom shows the physical length of chromosome 1 in Mb. **b**, Dot matrix plot showing the positions of homologous genes on chromosome 1. A BLASTP search was carried out using all of the predicted genes. A dot was plotted when the *E* value between two genes was less than 10^{-20} and the length of the match was over 50%. The colour of each dot represents the *E* value of the match. The colour spectrum bar at the bottom left shows the *E* value associated with each colour; for example, green corresponds to $E = 10^{-40}$. The matches of $E < 10^{-100}$ are shown as red dots. 'Self-against-self' matches are omitted.

There are at least three compelling reasons for obtaining finished high-quality sequence for the complete rice genome: first, the ability to determine gene function is highly dependent on having accurate sequences; second, as a model plant for the cereal grasses, the complete rice sequence will directly affect what can be accomplished with the other cereal grasses; and last, the identification of genes responsible for agronomic traits of economic importance requires precise map-based genomic sequence. Chromosome 1 contains many biologically important genes. More than 20 gene loci have been identified by genetic analysis, including genes controlling dwarfing and fertility. One of these genes, *sd1* has been cloned and shown to encode one of the enzymes in gibberellic acid synthesis[17].

The complete genomic sequence of chromosome 1 has yielded several findings that would be observed only using a clone-by-clone sequencing strategy. Gene families comprising active and inactive members and sets of tandemly repeated genes seem to be common features of chromosome 1. This redundancy may account for the unexpectedly large number of predicted genes on this chromosome. The intergenic repetitive fraction of the genome is not well understood and is frequently described as 'junk'. Repetitive sequences are usually removed or separated from other sequences before whole-genome shotgun assembly because they can cause global misassembly. But we know that functional genes are found in repetitive sequences and that transposable elements embedded in the repetitive sequences can restructure genomes, can control gene action and are likely to be involved in generating some of the allelic variation that has been selected in plants.

In addition, high-quality finished sequence provides the only real opportunity to study gene regulation, because most of the essential regulatory sequences fall outside the transcribed regions and our analysis of a restricted region of the genome showed that 43% of the genes predicted from whole-genome shotgun sequence methods were incomplete. Our results and those from the sequencing of rice chromosome 4 (ref. 18) show clearly the importance of the finished sequence. The IRGSP has an immediate goal of sequencing the rice genome to a minimum standard of the high-throughput genomic sequence (HTG) phase 2 level by the end of 2002 and is committed to a long-term goal of obtaining finished high-quality sequence for the whole genome. □

Methods

Chromosome sequencing

We sequenced the whole chromosome 1 of *Oryza sativa* ssp. *japonica*, variety Nipponbare, from 390 overlapping PAC/BAC clones. Initially, we constructed a sequence-ready physical map using the RGP *Sau*3AI PAC and *Mbo*I BAC libraries[19]. We also used *Hin*dIII or *Eco*RI BAC libraries constructed by Clemson University Genomics Institute (CUGI), and BAC clones with draft sequence data provided by Monsanto for gap filling in particular. We carried out shotgun sequencing of RGP and CUGI PAC/BAC clones to obtain sequence data with tenfold overlap. For Monsanto BAC clones[20], we complemented the available draft sequence (fivefold redundancy) with an additional fivefold overlap sequence (http://rgp.dna.affrc.go.jp/genomicdata/seqstrategy/newstrategy.html).

After the initial assembly of sequence data, stretches of poor or ambiguous quality and apparent gap regions were identified for further sequencing to obtain greater than 99.99% sequence accuracy. But despite extensive efforts to improve the sequence quality and to fill the gaps, 4 of the 390 PAC/BAC clones sequenced are still at phase 1 (GenBank, http://www.ncbi.nlm.nih.gov/HTGS/) because the consensus sequence could not be ordered correctly owing to numerous repeats. The remainder comprises 16 phase 2 and 370 phase 3 clones. The nine contigs for chromosome 1 representing the non-overlapping segments of continuous sequence were conjoined by inserting into the gap regions nucleotides that were calculated on the basis of the results of FISH experiments. All of the sequence information of chromosome 1 has been submitted to the DNA Data Bank of Japan (DDBJ, http://www.ddbj.nig.ac.jp/) with the accession number BA000010 (Con Division).

Gene prediction and functional classification

We carried out gene prediction using our in-house automated gene prediction system RiceGAAS[13]. The algorithm for gene domain prediction in RiceGAAS was designed by combining several prediction programs including GENSCAN[21] for maize, GENSCAN[21] for *Arabidopsis*, RiceHMM (http://rgp.dna.affrc.go.jp/RiceHMM/index.html) and the exon-finding program MZEF (http://argon.cshl.org/genefinder/), with homology search results from BLASTN and BLASTX (http://www.ncbi.nlm.nih.gov/BLAST/). These results were merged and integrated for gene prediction. Domain search was done using InterPro (http://www.ebi.ac.uk/interpro/scan.html), and repeats were identified using RepeatMasker (http://ftp.genome.washington.edu/cgi-bin/RepeatMasker). The predicted proteins were used to query the nonredundant protein database using BLASTP and categorized according to functional categories defined for *Arabidopsis* by MIPS (http://mips.gsf.de/cgi-bin/proj/thal/filter_funcat.pl?all) with a threshold probability value of 10^{-20}.

Received 4 April; accepted 19 September 2002; doi:10.1038/nature01184.

1. 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).
2. Wu, J. *et al.* A comprehensive rice transcript map containing 6591 expressed sequence tag sites. *Plant Cell* **14,** 525–535 (2002).
3. Yamamoto, K. & Sasaki, T. Large-scale EST sequencing in rice. *Plant Mol. Biol.* **35,** 135–144 (1997).
4. 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).
5. Yu, J. *et al.* A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science* **296,** 79–91 (2002).
6. Goff, S. *et al.* A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science* **296,** 92–100 (2002).
7. Cheng, Z. *et al.* Functional rice centromeres are marked by a satellite repeat and a centromere-specific retrotransposon. *Plant Cell* **14,** 1691–1704 (2002).
8. Dong, F. *et al.* Rice (*Oryza sativa*) centromeric regions consist of complex DNA. *Proc. Natl Acad. Sci. USA* **95,** 8135–8140 (1998).
9. Richards, E. J. & Ausubel, F. M. Isolation of a higher eukaryotic telomere from *Arabidopsis thaliana*. *Cell* **53,** 127–136 (1988).
10. The Arabidopsis Genome Initiative Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408,** 796–820 (2000).
11. Cheng, Z. *et al.* Toward a cytological characterization of the rice genome. *Genome Res.* **11,** 2133–2141 (2001).
12. Barakat, A., Carels, N. & Bernardi, G. The distribution of genes in the genomes of Gramineae. *Proc. Natl Acad. Sci. USA* **94,** 6857–6861 (1997).
13. Sakata, K. *et al.* RiceGAAS: an automated annotation system and database for rice genome sequence. *Nucleic Acids Res.* **30,** 98–102 (2002).
14. Hiratsuka, J. *et al.* The complete sequence of 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).
15. Lowe, T. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25,** 955–964 (1997).
16. Wessler, S. R., Bureau, T. E. & White, S. E. LTR-retrotransposons and MITEs: important players in the evolution of plant genomics. *Curr. Opin. Genet. Dev.* **5,** 814–821 (1995).
17. Sasaki, A. *et al.* A mutant gibberellin-synthesis gene in rice. *Nature* **416,** 701–702 (2002).
18. Feng, Q. *et al.* Sequence and analysis of rice chromosome 4. *Nature* **420,** 316–320 (2002).
19. 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,** 41–52 (2000).
20. Barry, G. The use of the Monsanto draft rice genome sequence in research. *Plant Physiol.* **125,** 1164–1165 (2001).
21. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268,** 78–94 (1997).

Supplementary Information accompanies the paper on *Nature*'s website (http://www.nature.com/nature).

Acknowledgements We thank Monsanto for the BAC contig information, BAC clones and their sequence data; R. Wing of Clemson University Genomics Institute and Novartis for the rice Nipponbare BAC library and its fingerprint data, respectively; M. Hattori for technical assistance; B. Burr and F. Burr for critically reading the manuscript; T. Slezak for comments; and K. Eguchi for encouragement.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to T. Sasaki (e-mail: tsasaki@nias.affrc.go.jp).

Sequence and analysis of rice chromosome 4

Qi Feng*†, Yujun Zhang*†, Pei Hao*†, Shengyue Wang†‡, Gang Fu‡, Yucheng Huang*, Ying Li*, Jingjie Zhu*, Yilei Liu*, Xin Hu*, Peixin Jia*, Yu Zhang*, Qiang Zhao*, Kai Ying*, Shuliang Yu*, Yesheng Tang*, Qijun Weng*, Lei Zhang*, Ying Lu*, Jie Mu*, Yiqi Lu*, Lei S. Zhang*, Zhen Yu*, Danlin Fan*, Xiaohui Liu*, Tingting Lu*, Can Li*, Yongrui Wu*, Tongguo Sun*, Haiyan Lei*, Tao Li*, Hao Hu*, Jianping Guan*, Mei Wu*, Runquan Zhang*, Bo Zhou*, Zehua Chen*, Ling Chen*, Zhaoqing Jin*, Rong Wang*, Haifeng Yin‡, Zhen Cai‡, Shuangxi Ren‡, Gang Lv‡, Wenyi Gu‡, Genfeng Zhu‡, Yuefeng Tu‡, Jia Jia‡, Yi Zhang‡, Jie Chen‡, Hui Kang‡, Xiaoyun Chen‡, Chunyan Shao‡, Yun Sun‡, Qiuping Hu‡, Xianglin Zhang‡, Wei Zhang‡, Lijun Wang‡, Chunwei Ding‡, Haihui Sheng‡, Jingli Gu‡, Shuting Chen‡, Lin Ni‡, Fenghua Zhu‡, Wei Chen§, Lefu Lan§, Ying Lai§, Zhukuan Cheng‖¶, Minghong Gu‖, Jiming Jiang¶, Jiayang Li§, Guofan Hong*, Yongbiao Xue§ & Bin Han*

** National Center for Gene Research, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, 500 Caobao Road, Shanghai 200233, China*
‡ Chinese National Human Genome Center at Shanghai, 351 Guo Shoujing Road, Shanghai 201203, China
§ Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Datun Road, Andingmenwai, Beijing 100101, China
‖ Yangzhou University, 27 Wenhua Road, Yangzhou, Jiangsu 225009, China
¶ Department of Horticulture, University of Wisconsin, Madison, Wisconsin 53706, USA
† These authors contributed equally to this work

Rice is the principal food for over half of the population of the world. With its genome size of 430 megabase pairs (Mb), the cultivated rice species *Oryza sativa* is a model plant for genome research[1]. Here we report the sequence analysis of chromosome 4 of *O. sativa*, one of the first two rice chromosomes to be sequenced completely[2]. The finished sequence spans 34.6 Mb and represents 97.3% of the chromosome. In addition, we report the longest known sequence for a plant centromere, a completely sequenced contig of 1.16 Mb corresponding to the centromeric region of chromosome 4. We predict 4,658 protein coding genes and 70 transfer RNA genes. A total of 1,681 predicted genes match available unique rice expressed sequence tags. Transposable elements have a pronounced bias towards the euchromatic regions, indicating a close correlation of their distributions to genes along the chromosome. Comparative genome analysis between cultivated rice subspecies shows that there is an overall syntenic relationship between the chromosomes and divergence at the level of single-nucleotide polymorphisms and insertions and deletions. By contrast, there is little conservation in gene order between rice and *Arabidopsis*.

The rice genome has been well mapped both genetically and physically[3–5] and has a syntenic relationship with other cereals[6]. *Arabidopsis thaliana* (Arabidopsis), a member of the Brassica family of dicotyledonous (dicot) plants, has become an important model flowering plant for studying many aspects of plant biology[7]. The completion of the Arabidopsis genome[8–10] has afforded an unprecedented opportunity for systematic studies of plant gene function. Equally, the complete rice genome sequence will provide a catalogue of genes that may be important for improving not only rice but also other cereals, as functionally important sequences are conserved and may be identified by their similarity[11]. The International Rice Genome Sequencing Project (IRGSP) has adopted the clone-by-clone approach for obtaining a finished rice genome sequence, because it is modular, allows efficient gap filling, avoids problems arising from distant repetitive sequences and results in the early completion of larger contiguous segments of a genome. Here we describe the completed sequence of rice *O. sativa* ssp. *japonica* cv.

